1 Recognizing Between-Person Differences and Within-Person Fluctuations in Affect Experience from Speech Collected with Smartphones

1.1 Abstract

Speech data have become ubiquitous and are analyzed by emotion-detecting AI services. However, those algorithms that have been trained on enacted or labelled speech samples representing affect *expression* are used to infer subjective affect *experience*. In the present study, we investigated if machine learning algorithms can recognize subjective affect experience from speech samples collected in the wild. In two studies, we extracted acoustic voice parameters and state-of-the-art word embeddings from 19,101 speech samples with corresponding experience-sampled self-reports on affect experience from 1,071 participants collected using smartphones. We showed that affect experience on the dimension of overall arousal and contentedness, but not overall valence and sadness were significantly predictable using machine learning algorithms. Further, experimental and empirical findings suggest that emotional speech content did not affect predictions from voice acoustics (i.e., what someone talks about does not affect how well emotions can be predicted from voice cues alone). We discuss implications for the algorithmic monitoring of affect experience and raise issues regarding the protection of user privacy rights, for example in the context of voice assistants.

1.2 Introduction

Advances in the algorithmic recognition of affective states and related affective disorders from speech offer promising applications, for instance in marketing and health care (Hildebrand et al., 2020; Milling, Pokorny, Bartl-Pokorny, & Schuller, 2022). Prior research has successfully predicted affective states from speech cues in a range of data sources, such as TV clips, phone calls, and enacted speech Vogt, André, & Wagner (2008). The promises of algorithmic affect recognition from speech and the ubiquity of speech data due to the rise of voice assistants, for example Amazon's Alexa and Apple's Siri, have also created an increasing commercial interest in the field. Tech companies aim to leverage those speech data to, for instance, recognize what affect their customers experience (Mandell, 2020; **knightAmazonWorkingMaking2016a?**) on order to develop personalized user interfaces, for example by mimicking the user's characteristics (Vlahos, 2019). Most of the research in automatic affect recognition from speech and the corresponding commercial tools, for example in the area of mental health care, are trained on enacted or labelled speech samples that represent emotion *expressions*. However, those algorithms are used to detect people's subjective affect *experience*. Further, many of the commercial algorithms are not transparent with regard to how well their predictions work and how predictions are being made. This raises issues regarding the promises of emotion recognition speech technology and the protection of user privacy in setting where speech data can be analyzed, for example, when using voice assistants. The present work investigates the algorithmic prediction of between-person differences and within-person fluctuations in subjective affect experience from speech samples collected with smartphones.

1.2.1 Predicting Affect from Speech

Researchers have successfully predicted affect from a range of speech data, such as labelled TV clips (Grimm, Kroschel, & Narayanan, 2008) and enacted speech samples from the lab (Bänziger, Mortillaro, & Scherer, 2012; Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). They report impressive prediction performance for the automatic recognition of emotions (i.e., correlations between true scores and predicted scores of up to .81 for arousal and .68 for valence predictions) (Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013). However, one has to keep in mind that in those works the enacted target emotion or the rater labels serve as ground truth. Moreover, the prediction performance varies greatly across studies due to a varying choice of emotion targets (i.e., discrete emotion vs. core affect), conceptualizations of affect (e.g., short-termed elicited emotions vs. moods), and employed algorithms (e.g., supervised vs. unsupervised machine learning).

These studies on algorithmic affect recognition, however, often offer no insights into how predictions in their "black-box" models were being made. For instance, it frequently remains unclear which specific speech characteristics are particularly predictive of an affective state. Prior descriptive research reported on associations of specific acoustic features and affective states. For example, pitch and intensity were found to be associated with arousal (Vogt, André, & Wagner, 2008; Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013). Two recent studies provide a remarkable non-technical summary of voice features (Hildebrand et al., 2020) and a comprehensive overview of associations of word dictionaries with affect in spoken language (Sun, Schwartz, Son, Kern, & Vazire, 2020). Recent developments in the area of interpretable machine learning can help to gain insights into the inner working of machine learning algorithms and, consequently, aid with detecting speech features are predictive of affective states (Molnar, 2019).

Due to the challenge of obtaining speech data wit corresponding affect labels in-vivo, most prior research on affect recognition from speech has used actors or labelled samples. This comes with a set of downsides, such as actors potentially overacting and the ambiguity of ground truth due to the subjective nature of labeling (see chapter 1) (Batliner et al., 2011; Schuller, 2018; Wilting, Krahmer, & Swerts, 2006). As a consequence, studies investigating predictions of subjective affect experience from speech are rare. Recent works have collected everyday speech samples using the Electronically Activated Recorder (EAR) (Mehl, 2017). Hereby, more data can be collected over a period of time which allows researchers to not only investigate between-person differences in affect (i.e., is this person sad?), but also assess within-person fluctuations (Huang & Epps, 2018; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). Using the EAR, however, can be privacy invasive since potentially non-consenting persons may be recorded, too. Moreover, handling the EAR recorders and transmitting the collected data can be tedious for participants and researchers. Off-the-

shelf smartphones represent a useful platform to collect experience samples on momentary affect experience over time and make corresponding speech records using the build-in microphone (Weidman et al., 2020). Recent works leveraging the EAR and smartphones have reported low prediction performance for subjective affect experience from speech collected in the wild (Carlier et al., 2022; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020).

Prior research has shown that voice acoustics (prosody) and the lexical content of the produced words (semantics) work together when transmitting affective information through speech (Ben-David Boaz M., Multani Namita, Shakuf Vered, Rudzicz Frank, & van Lieshout Pascal H. H. M., 2016). Moreover, studies suggest that there is a prosodic dominance in the perception of affect (Ben-David Boaz M., Multani Namita, Shakuf Vered, Rudzicz Frank, & van Lieshout Pascal H. H. M., 2016; Lin Yi, Ding Hongwei, & Zhang Yang, 2020) based on lab experiments, but not (yet) using language data from the wild (Schwartz & Pell, 2012). Moreover, while this research filed focused on the interplay of prosody and semantics in the recognition of affect by human raters, there are, to our knowledge, no studies on affect-prosody interactions in algorithmic affect detection. Hence, it is unclear if what users talk about (i.e., the emotional content) has an effect on voice acoustics that impacts automated affect recognition. In applied setting, for example, the question is if an algorithm could an recognize affective state regardless of what the person talks about, may it be a mundane topic, such as the weather or ordering pizza, or does one need to talk about an emotional topic (e.g., meeting a loved one).

The present work leverages methodological advances in the area of smartphone-based data collection methods to investigate the prediction of between-person differences and within-person fluctuations in subjective momentary affect experience from speech. In two large-scale studies, we train cross-validated machine learning models on acoustic voice cues and state-of-the-art word embeddings from speech samples collected in the wild. Moreover, for predictive models we investigate which speech cues were most predictive for affect experience. Further, we experimentally and empirically investigate the effects of speech content on algorithmic affect recognition from voice acoustics. Thereby, we aim to advance potential applications and promises in automatic affect recognition from speech signals, and inform the discussion on the protection of user privacy rights.

1.3 Study 1.1

1.3.1 Method

1.3.1.1 Smartphone-Based Voice Data Collection and Privacy-Preserving On-Device Acoustic Feature Extraction Data collection for this study was part of a large six-month panel study (from May until November 2020) using the *PhoneStudy* research app at Ludwig-Maximilian-Universität München (Schoedel & Oldemeier, 2020). Data collection was approved by the LMU IRB board and data privacy office. The study comprised two two-week experience sampling phases (July 27, 2020 to August 9,2020; September 21, 2020 to October 4, 2020) during which participants received two to four short questionnaires per day. Here, self-reported valence and arousal were assessed in two separate items on six-point Likert scales among other psychological properties as part of an experience sampling procedure. Furthermore, for each experience sampling instance, we computed the fluctuation of assessed momentary affect in valence and arousal from one's (median) affect baseline (for participants with at the five experience sampling days) across all experience sampling instances. For example, if a participant had a valence baseline of "3" and reports a "6" in a particular moment, this fluctuation score of "+3" indicated that this person had been a lot more happy than usual.

The last experience sampling questionnaire of each day included an additional instruction, where participants were asked to read out a series of predefined emotional 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 affective sentences (Defren et al., 2018) and differ in their emotional content: positive (e.g., "My team won yesterday."), negative (e.g., "Nobody is interested in my life."), and neutral (e.g., "The plate is on the round table."). These three emotional categories are presented consecutively in each audio logging task. The order of the categories was randomized per experience sampling questionnaire. For each emotional content category three sentences were randomly drawn from respective sets of sentences in the database. The use

and experimental manipulation of these semantic categories allowed us to control the content spoken by our participants, but at the same time allowed us to conduct a privacy-friendly study. The audio recording was started 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. We chose these lower and upper time thresholds because this is the minimum and maximum time required to read out the three sentences. Once the audio record had been completed, we automatically extracted voice parameters using the widely adopted open source OpenSMILE algorithm (Eyben, Wöllmer, & Schuller, 2010) to extract two sets of acoustic features directly on the participant's device: First, the extended Geneva Minimalistic Acoustic Parameters Set (eGeMAPS) comprised of 88 features (Eyben et al., 2016). The feature sets are clustered into feature groups termed low level descriptors (LLD) (e.g., Loudness, Pitch, Frequency). Theses feature sets have been used in prior studies on affect recognition from acted or labelled voice samples. After feature extraction the voice records were automatically deleted and only extracted voice features were stored on our servers.

With this procedure, we collected 11217 audio logs from 587 participants. We excluded 215voice logs because the respective acoustic features (Voiced segments per second) indicated that no human voice was recorded. Moreover, we excluded samples without corresponding self-reports on valence and arousal, participants with less than five experience sampling days, and those participants that had no variance in all of their valence and arousal scores across all their experience samples. This left us with a final data set of 9922 voice samples with corresponding acoustic features from 3381 experience sampling instances for valence and arousal from 499 participants (48.4 % female, M(Age) = 42.97 years). Overall self-reported valence was positive (M =4.72, SD = 1.03) and overall arousal was slightly geared towards activity (M =3.68, SD = 1.35). The distribution of valence and arousal as well as fluctuations from the baseline is provided in the appendix.

In the final sample, voice samples were equally distributed across conditions (2(2) =37.4, p > .05): 3224 came from the positive condition, 3111 from the neutral condition, and 3587 from the negative condition.

1.3.2 Predictive Modelling

We trained multiple supervised machine learning regression models on the extracted acoustic features for the prediction of self-reported valence and arousal. Machine learning algorithms have multiple advantages over classical regression and have been used in prior research (Weidman et al., 2020). Here, we compared the predictive performance of LASSO regularized regression models (**zouRegularizationVariableSelection2005?**) with those of a non-linear tree-based random forest model (Wright & Ziegler, 2017; **breimanRandomForests2001?**), and a baseline model. The baseline model would predict the respective mean values for valence and arousal of the respective training set for all cases in a test set. We evaluated the predictive performance of our models in a cross-validation scheme (**bischlResamplingMethodsMetaModel2012?**). Additionally, we included the prediction of age as a benchmark. We blocked 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 evaluated the predictive performance of the models based on the coefficient of determination (R^2) and Spearman (rank) correlation (r) and between the predicted scores and participants' self-reported scores. To determine whether a model was predictive (alpha = 0.05) at all, we carried out t tests (one-sided) by comparing the R^2 measures in all prediction models with those of the baseline models. We used variance corrected t tests based on 10-fold cross- validation to account for the dependence structure of cross-validation experiments (**nadeauInferenceGeneralizationError2003?**). All comparisons were adjusted for multiple comparisons (n = 2) via Holm correction.

All data processing and statistical analyses in this work were performed with the statistical software R version 4.1.1 (**rcoreteamLanguageEnvironmentStatistical2021?**). For machine learning, we used the mlr3 framework (Lang et al., 2019) and the DALEX package (Biecek, 2018). We preregistered the present study as a transparent account of our work (Koch & Schoedel, 2021).

1.4 Results

1.4.1 Recognizing Affect Experience from Acoustic Voice Cues

Overall, none of the employed algorithms predicted affect experience significantly better than chance, even though, on average, predictions of arousal fluctuations from Random Forest models (R^2 = 0, r = 0.11) were better than the baseline models (R^2 = -0.01, r = -0.01, r = NA). Overall, even though not significantly better than chance, prediction models' performance was better for raw arousal and fluctuation than for valence. Figure @ref(fig:ger prediction overview) provides an overview of the performance of all learners across prediction tasks. On the contrary, predictions of speaker age (R^2 = 0.11, r = 0.38) and gender (prediction accuracy = 91.31 %) suggest that voice acoustics from read-out sentences contain information about speaker demographics.

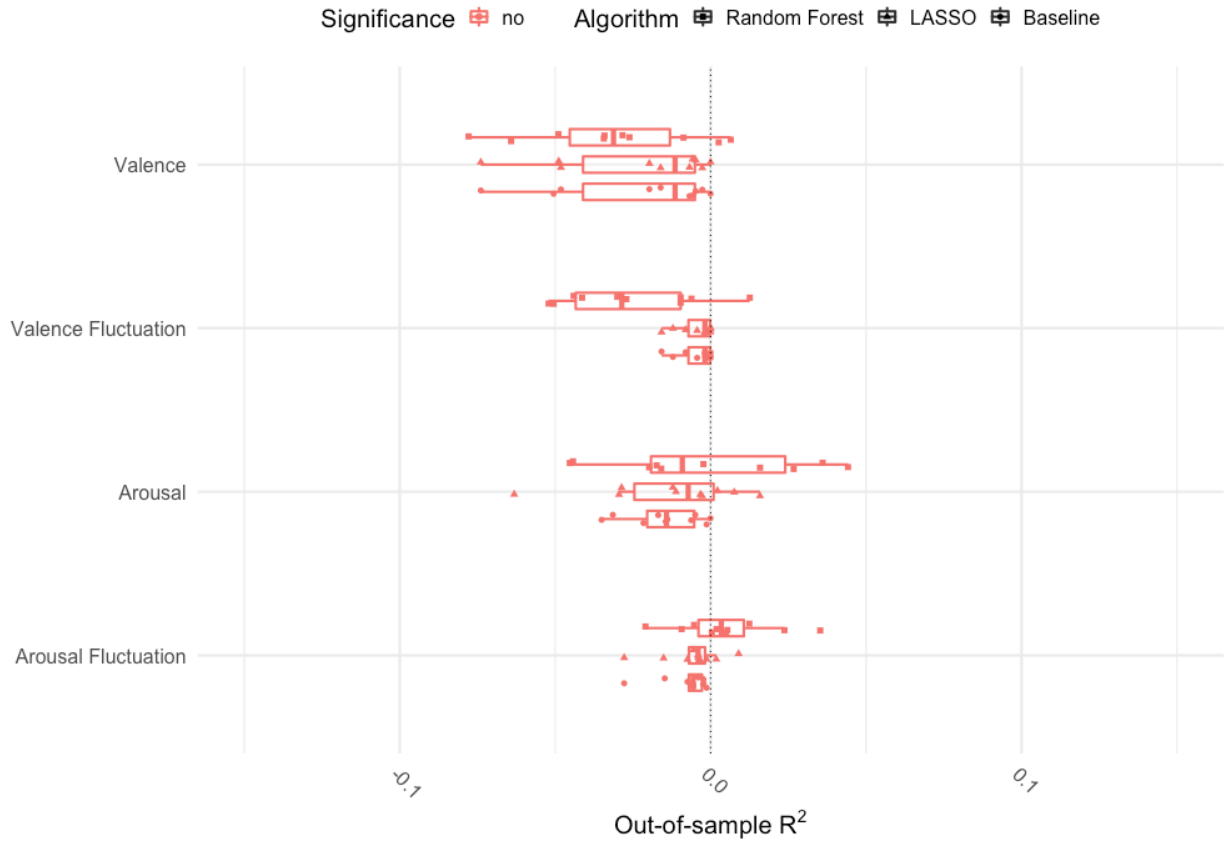


Figure 1: Box and whisker plot of prediction performance measures from 10-fold cross-validation for prediction models.

Features related to spectral flux and loudness were most important. This is in line with descriptive correlations of voice features and affect experience (see appendix).

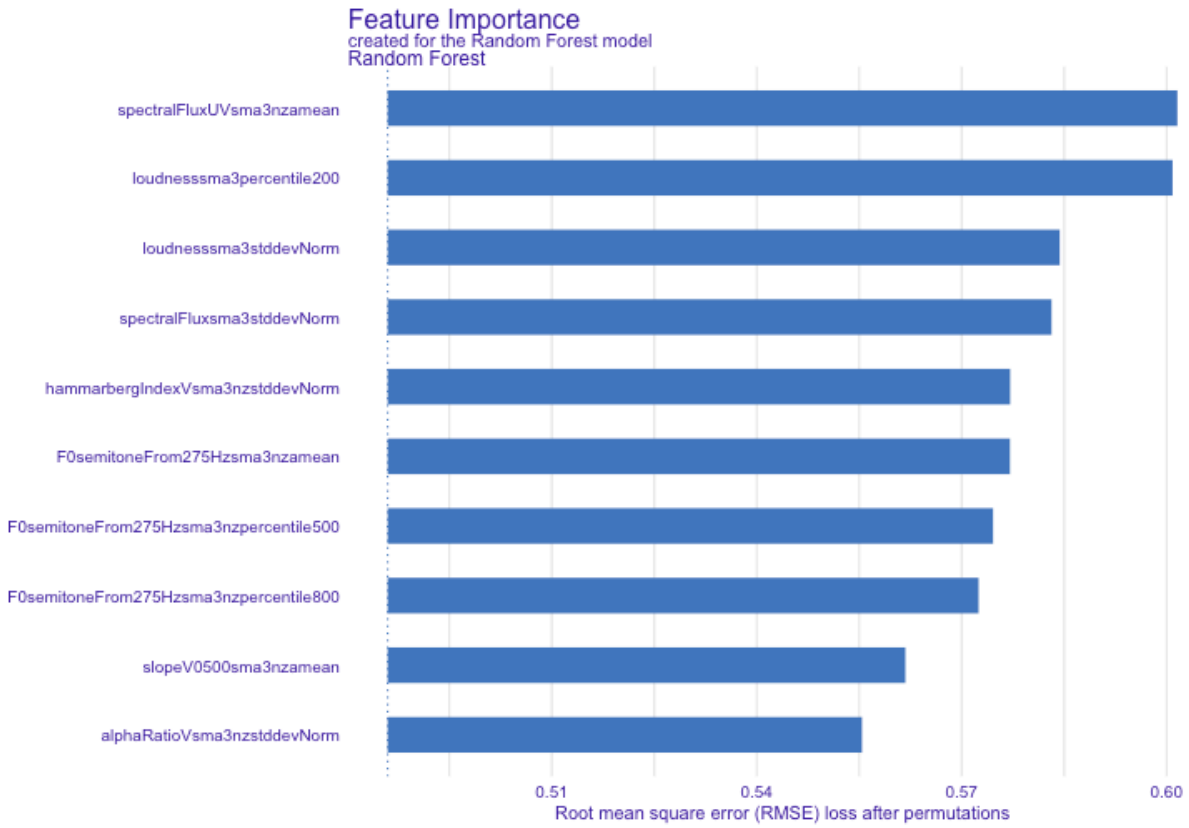


Figure 2: Permutation feature importance for the most predictive features in the Random Forest model for arousal prediction. Permutation feature importance represents the decrease in the model's prediction performance (as measured by RMSE) after permuting a single variable.

1.4.2 Content Effects on Affect Predictions from Voice Acoustics

Finally, we analyzed if the experimentally altered emotional content (positive/ neutral/ negative) of the predefined sentences that had been read out by participants an effect on affect predictions from voice acoustics. Since between-person differences showed better, but still not significant, prediction performance than within-person fluctuations, we focused on the former. Here, our results indicate that sentences' emotional valence did not matter. Moreover, there were no differences in prediction errors for valence and arousal predictions (see Figure @ref(fig:ger content effect)).

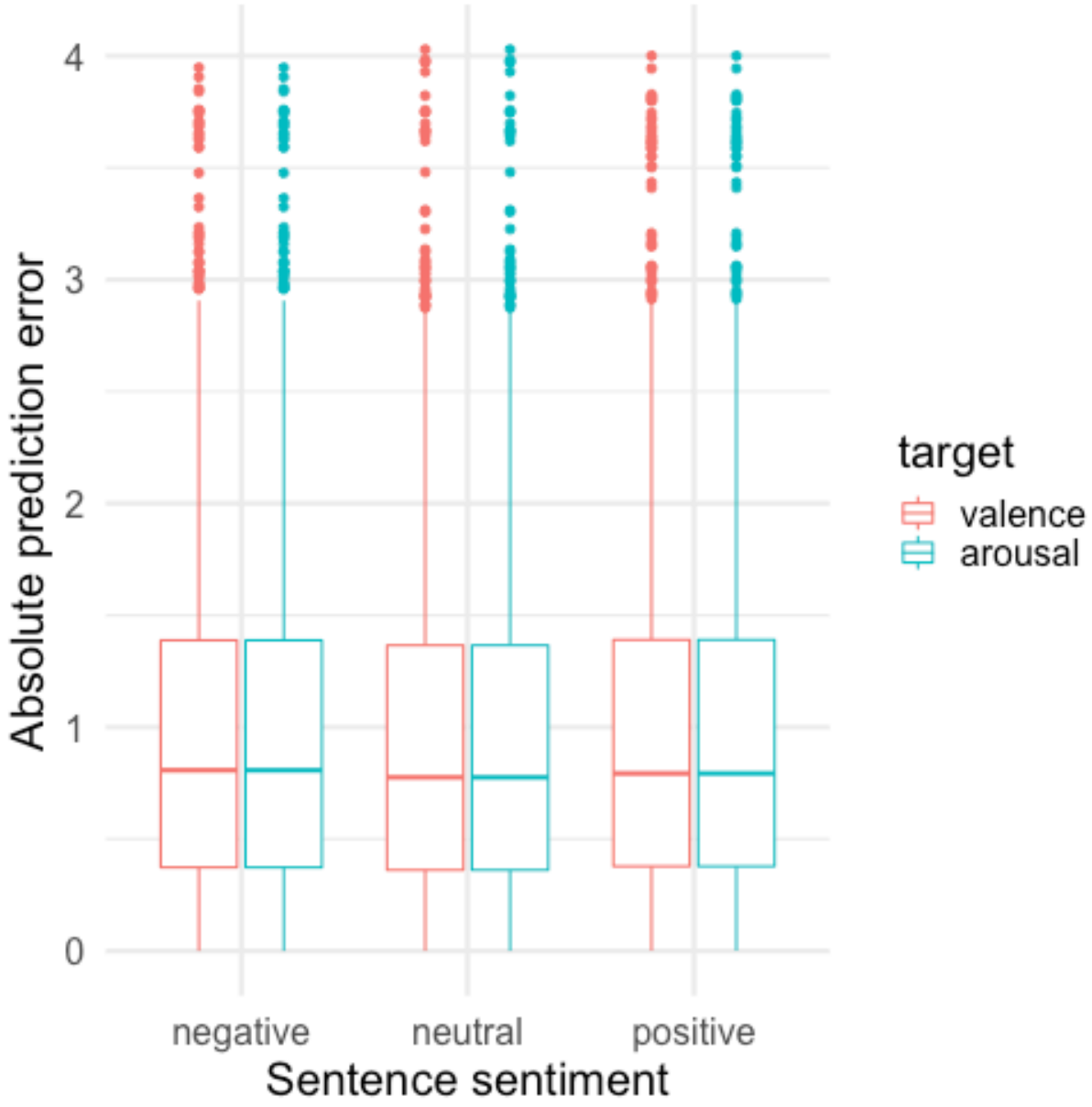


Figure 3: Prediction error in different emotional sentence conditions.

1.5 Study 2

1.5.1 Method

1.5.1.1 Smartphone-Based Speech Data Collection Data collection for this study was part of UT1000 Project at the University of Texas at Austin in the United States in Fall 2018(Wu et al., 2021). During a three-week self tracking assignment using their smartphones, students from a Synchronous Massive Online Course (SMOC) version of Introductory Psychology received four short experience sampling questionnaires per day where they could also make records of their speech at the end. Here, self-reported arousal, contentedness, and sadness were assessed in separate items on four-point Likert scales among other psychological properties as part of an experience sampling procedure. Thereby, in the second study, we assessed emotional valence on two items instead of one as in the first study. According to the affect grid, contented-

ness and sadness have the same low level of arousal and an opposing emotional valence. Furthermore, for each experience sampling instance, we computed the fluctuation of assessed momentary affect in arousal, contentedness, and sadness from one's (median) affect baseline (for participants with at least five experience sampling days) across all experience sampling instances. For the audio records, participants received the following instruction: "Please record a short audio clip in which you describe the situation you are in, and your current thoughts and feelings. Collect about 10 seconds of environmental sound after the description." The reposes to this prompt had been analyzed in the present study. Any parts of the record that did not contain any speech were cut out later since the focus of this work had been affect in human speech. The speech data had also been used in another project that describes the data collection procedure more in detail (Marrero, Gosling, Pennebaker, & Harari, 2022).

With this procedure, we collected 23,482 audio logs from 980 participants. We excluded speech logs they contained too little speech (i.e., less than 15 spoken words and four seconds of recorded speech) or because the respective acoustic features (Voiced segments per second) indicated that no human voice was recorded. Moreover, we excluded samples without corresponding self-reports, participants with less than five experience sampling days, and those participants that had no variance in all of their valence and arousal scores across all their experience samples. This left us with a final data set of 13724 speech samples with corresponding experience-sampled self-reports on momentary affect experience from 567 participants (35.1 % female, M(Age) = 18.57 years). Overall participants reported positive level of experienced contentedness (M =1.65, SD = 0.85) and low sadness (M =0.53, SD = 0.77). Overall arousal was balanced out (M =1.95, SD = 0.95). The distribution of arousal, contentedness, and sadness as well as respective fluctuations from the baseline is provided in the appendix.

In the same manner as in study 1.1, we extracted acoustic voice feature (eGeMAPS) audio samples using OpenSMILE from the raw audio files (Eyben et al., 2016; Eyben, Wöllmer, & Schuller, 2010; Schuller et al., 2016). In study 1.2, those features were extracted on the raw audio files after data collection and not directly on the smartphones as in the first study.

We transcribed all raw audio records using the Google Speech-to-text API. To ensure comparability of the two studies with regard to the length of speech samples, we retained all speech transcripts that contained at least 15 words and were more than four seconds long which is equivalent to the length of the sentences that had been read out in the first study. Then, we extracted state-of-the-art word embeddings from speech transcripts using the text R package (Kjell, Giorgi, & Schwartz, 2021). Specifically, for predictive modeling, we used the second to last layer (layer 23) from "RoBERTa large" as recommended in prior work (Liu et al., 2019; Matero, Hung, & Schwartz, 2022).

1.5.1.2 Predictive Modelling

For predictive modelling, we applied the same machine learning pipeline as used in study 1.1 to predict self-reported sadness, contentedness, and arousal as well as their deviation from the respective person's baseline level. Moreover, in addition to extracted acoustic features, we also used word embeddings as features. To investigate the predictive power of each feature set, we ran predictions on all features combined as well as acoustic features and word embeddings only.

1.5.2 Results

1.5.2.1 Prediction of Affect Experience from Speech

Our employed machine learning models trained on voice acoustics (speech form) and word embeddings (speech content) predicted between- person differences and within-person fluctuations in the subjective experience of momentary affect experience. Figure @ref(fig:us prediction overview) provides an overview of the performance of all learners across prediction tasks. Specifically, they yielded the best prediction performance for between-person variations in arousal (R^2 = 0, r = 0.12), contentedness (R^2 = 0, r = 0.12), and sadness (R^2 = 0, r = 0.12). Also, for within-person fluctuations, predictions were significantly better than chance for arousal (R^2 = 0, r = 0.12), contentedness (R^2 = 0, r = 0.12), and sadness (R^2 = 0, r = 0.12). However, overall, predictions were better for between-person differences than for within-person fluctuations.

Moreover,evaluation prediction performance of models trained only on voice acoustics or word embedding respectively revealed that predictions were mostly driven by the information coming from word embeddings.

While voice acoustics only provided significant affective information on contentedness (R^2 = 0, r = 0.12) and arousal (R^2 = 0, r = 0.12), word embeddings yielded much stronger predictions. Here, word embeddings trained on the content of the audio samples were significantly predictive of all targets: Between-person differences in arousal (R^2 = 0, r = 0.12), contentedness (R^2 = 0, r = 0.12), and sadness (R^2 = 0, r = 0.12) as well as within-person fluctuations of arousal (R^2 = 0, r = 0.12), contentedness (R^2 = 0, r = 0.12), and sadness (R^2 = 0, r = 0.12).

Finally, while for voice acoustics, as in the first study, Random Forest model performed better, suggesting non-linear relationships, for word embeddings LASSO models performed better, indicating linear predictor-outcome relationships.

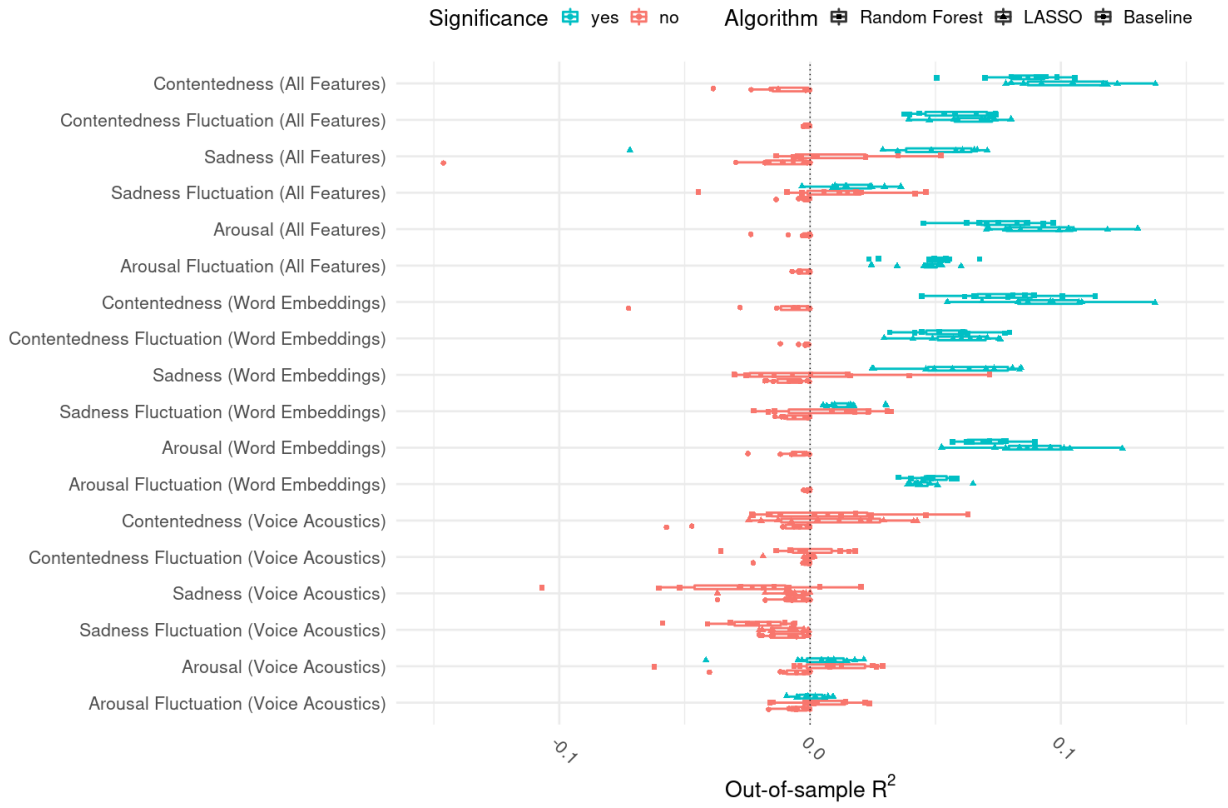


Figure 4: Performance of algorithms in different predictions tasks 10-fold cross-validation for each feature (sub) set.

Features related to spectral flux and loudness were most important. This is in line with descriptive correlations of voice features and affect experience (see appendix).

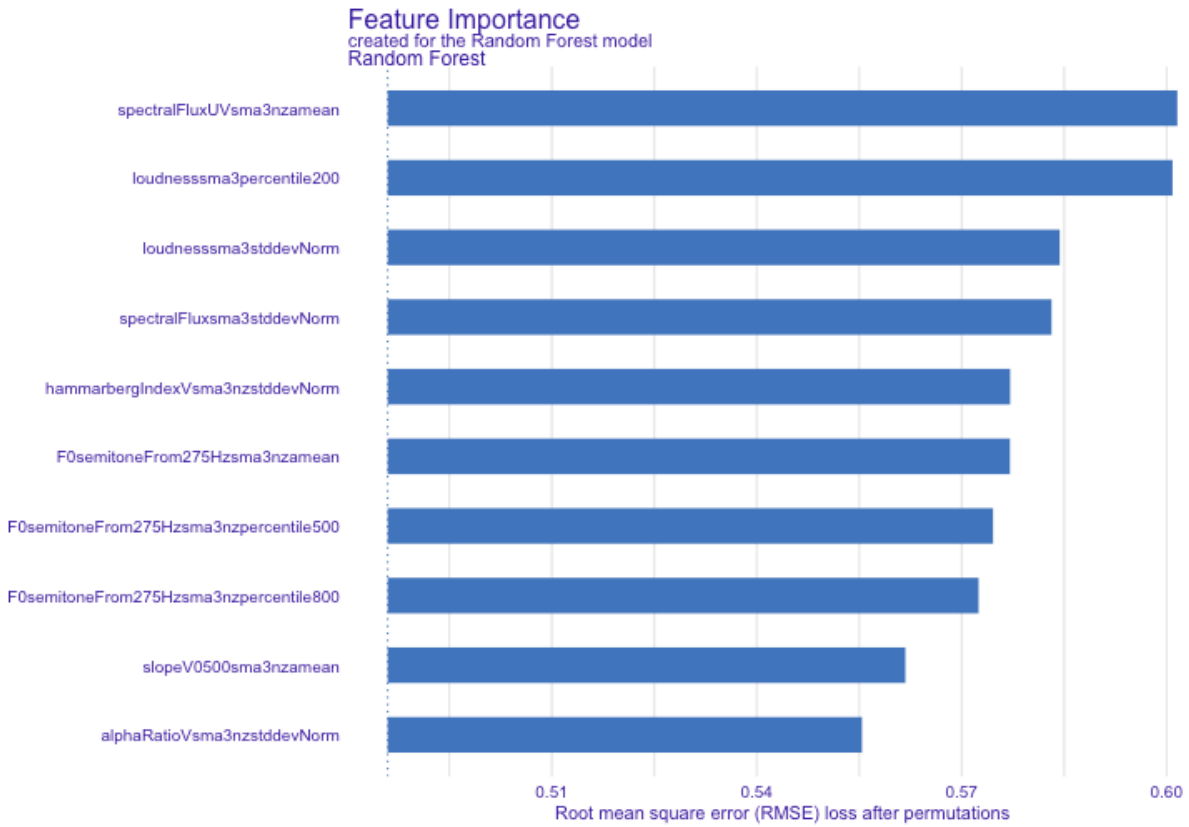


Figure 5: Permutation feature importance for the most predictive features in the Random Forest model for arousal prediction. Permutation feature importance represents the decrease in the model's prediction performance (as measured by RMSE) after permuting a single variable.

1.5.2.2 Content Effects on Affect Predictions from Voice Acoustics In order to empirically investigate the effect of the emotional valence of the spoken content on affect predictions from voice cues, we used the sentiment score (M =0.02, SD = 0.29) within the interval of [-1; 1] that had been assigned to each speech transcript by the Google text-to-speech API and analyzed the absolute prediction errors in arousal, contentedness, and sadness predictions form voice cues. Like in the first study, since between-person differences showed superior prediction performances than within-person fluctuations, we focused on the former. Figure @ref(fig:us content effect) depicts the audio sample's sentiment score on the x-axis and the absolute prediction error on the y-axis. Result indicate that while content sentiment did not have a clear effect on arousal and sadness predictions, extreme content sentiment seems to impoverish affect recognition of contentedness from voice cues (especially when content sentiment is contrary to prediction target; e.g., when predicting contentedness and content is very negative. However, absolute differences in prediction error with sentiment were small overall.

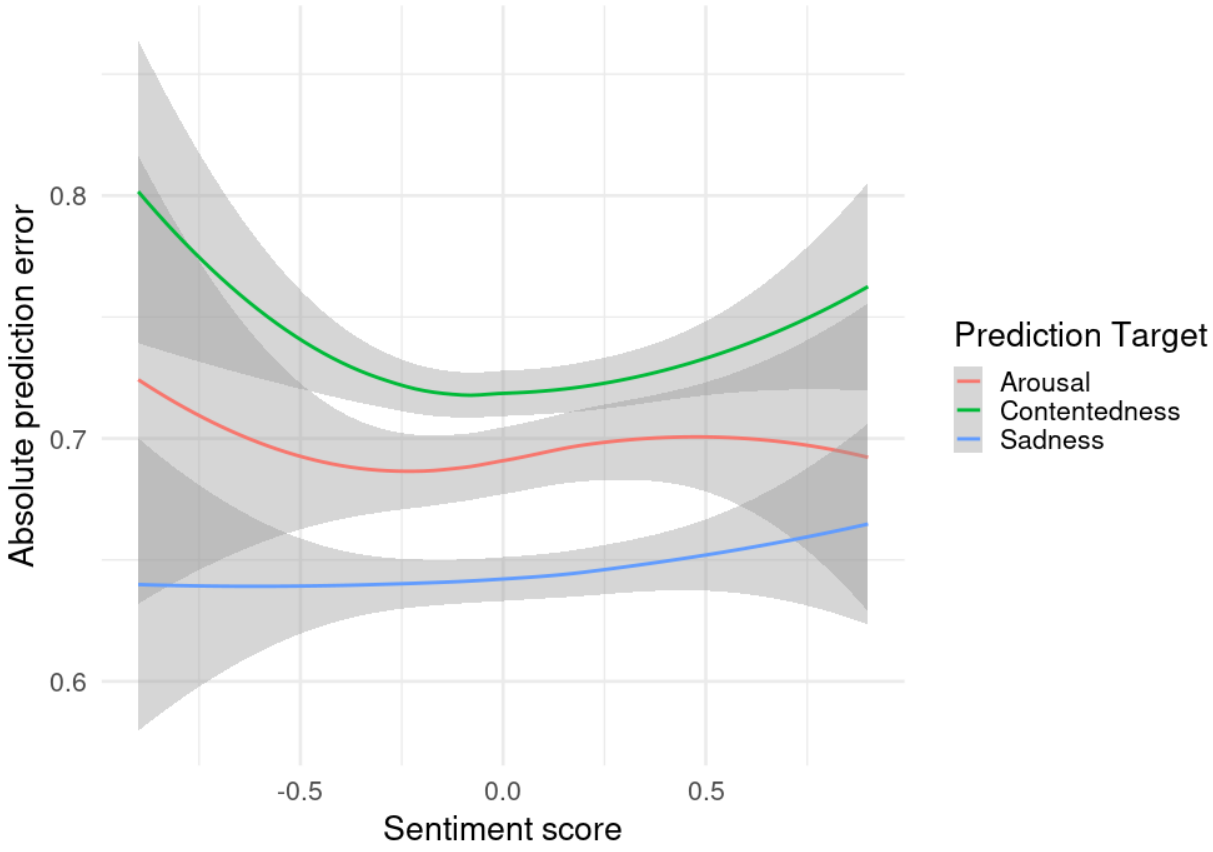


Figure 6: Sentiment score of content plotted against the absolute prediction error from acoustic predictions

1.6 General Discussion

In the present work, we extracted acoustic voice parameters and state-of-the-art word embeddings from speech samples collected using smartphones to predict between-person difference and within-person fluctuations in affect experience. While we assessed core affect on the dimensions of valence and arousal in a representative sample from Germany, in the second study based on US data, we used distinct variables instead of dimensional approach. Our models predicted overall arousal and contentedness, but not overall valence and sadness significantly better than chance. Overall, predictions were better when participants could talk freely (vs. reading out loud predefined emotional sentences). In our models, speech content showed superior prediction performance compared to voice acoustics and we identified loudness and frequency to be particularly predictive voice characteristics. Experimental (study 1.1) and empirical (study 1.2) findings suggest that emotional speech content did not affect predictions from voice acoustics (i.e., what someone talks about does not affect how well affect can be predicted from voice cues alone).

1.6.1 Recognizing Affect Experience from Speech Cues

Our results indicate that momentary affect experience of overall arousal and contentedness, but not overall valence and sadness can be automatically recognized from speech. Our machine learning models achieve lower prediction performance than prior work on automatic prediction of affect expression (Schuller, 2018) and similar on affect experience (Carlier et al., 2022; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). Prior works suggest that real-life affect is more difficult to recognized that enacted one (Vogt, André, & Wagner, 2008). Also, there are less instances of extreme affect experiences in our data set compared to the data used in prior studies on acted emotions. We rather predicted mood, which is, by definition, less intense than emotions, which are short-lived and directed (Scherer, 2003). Across the two studies, arousal

predictions were better than those for emotional valence, underlining prior work showing that the latter is more challenging to automatically infer due to its individual nature (Sridhar & Busso, 2022). Moreover, overall predictions of between-person differences were better than within-person fluctuations. This finding is in line with prior studies on the prediction of subjective affect experience from speech (Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). Thereby, our findings challenge the optimistic results on affect recognition from prior research from have broad implications for the monitoring of affect experience in a practical setting. They question the proclaimed performance of commercial affect recognition algorithms and highlight the challenges ahead in affect monitoring, for example in mental health care. Current expectations, particularly of commercial solutions for affect predictions from speech, might be overoptimistic. Finally, still this has implications for the privacy of users of, for example, voice assistants. In future research, smartphone could play a prominent role to collect and analyze speech data with corresponding in-situ self-reports on subjective affect experience for affect inferences. Hereby, smartphones could be used as a mobile experimental lab (Miller, 2012). Further, in line with prior research (Weidman et al., 2020), our results suggest that an economic acoustic feature set is sufficient for affect detection from voice. Small features set is less computationally expensive and would allow for online or on-device pre-processing in a scientific or applied setting.

We also compared the prediction performance of machine learning models trained on the much larger Compare2016 (6,737 features) feature set instead of the economic eGeMaPs (88 features) set, but more acoustic features did not yield better results. This is relevant for potential application of affect recognition on device.

1.6.2 The Context of Speech Production Matters

Our results show, that the context of speech production (i.e., fixed sentences vs. semi-structured prompt) had an impact on affect predictions. Our findings suggest that Study 2 predictions overall and for voice acoustics only might be better because participants could talk freely. While the predefined sentences allowed us to control for the semantics of what participants talked about in their voice records we let them record, they we unable to express themselves freely. As a result, researchers and practitioners should consider the context in which speech had been produced in and keep in mind that findings and trained models might be specific to the given production context and do no necessarily generalize well to other production contexts.

1.6.3 The Role of Speech Content

In study 2, state-of-the-art word embeddings show a superior affect prediction performance compared to voice acoustics suggesting that speech content could contain more affective information than speech form. This finding is in line with prior research that found speech content to be more predictive than voice acoustics when predicting momentary subjective experience of happiness (Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). As a consequence, event though prior research has suggested that voice acoustics could be more relevant for human affect inferences than speech content (Ben-David Boaz M., Multani Namita, Shakuf Vered, Rudzicz Frank, & van Lieshout Pascal H. H. M., 2016; Lin Yi, Ding Hongwei, & Zhang Yang, 2020), future research and AI application should consider both channels - content and form - simultaneously.

By experimentally varying the emotional valence of the spoken content in study 1 and empirically investigating the effect of word sentiment in study 2, our findings suggest that the content about what participants talked about, did not have a substantial impact on the algorithmic affect recognition from voice cues. This insight could imply that it does not matter what people talk about when algorithmically inferring affect experience from voice cues. However, one has to keep than in mind that affect predictions from voice only were overall not very strong, particularly in study 1. More research is needed to disentangle speech content and form in automatic affect recognition.

1.6.4 Affect-Linked Voice Variations

In our (out-of-sample) predictions, voice cues related to loudness and spectral flux were most relevant. These insights are in line with descriptive in-sample correlations of voice features and affect self-reports (see appendix). This observation is in line with prior work (Hildebrand et al., 2020).

1.6.5 Limitations and Future Directions

In this section, we discuss the two specific limitations of this study. General limitations related to the data collection method and the measurement of self-reported affect experience that are discussed in the general discussion (see chapter 4).

First, we used slightly different operationalizations of affect experience and sample compositions in the two studies that might affect their comparability: In study 1, we assessed valence and arousal on a six-point Likert scale. In study 2, we used two items to assess affective valence (contentedness and sadness) and arousal on a four-point Likert scale. As a consequence, findings are not directly comparable. Further, while study 1 drew on a representative German sample, study 2 was based on a student convenience sample from the United States with the respective limitations, such as potential limitations in generalizability of findings (Müller, Chen, Peters, Chaintreau, & Matz, 2021). Future studies should investigate multiple target emotions in diverse international samples from different cultural contexts and non-Western countries.

Second, and most importantly, in contrast to prior work using passive speech sensing (e.g., via EAR), participants had to actively log their speech. This artificial setting might have an effect on results. Moreover, the findings of this study might be subject to the specific instructions that had been given for the audio records: In study 1, participants were instructed to read out predefined sentences and, in study 2, participants were prompted to talk about the situation they were in as well as their current thoughts and feelings in a semi-structured fashion. While affect-linked acoustic voice cues in the two studies look similar and are most likely transferable to new voice data, word embeddings are specific to the given task in study 1.2. In this manner, future work should employ multiple different speech tasks for affect predictions and investigate how well predictions generalize from one prompt to another. Moreover, specifically to study 1.1, another related limitation lies in our privacy-preserving on-device pre-processing approach. By applying on-device feature extraction, we had no opportunity to check in detail if participants truly complied with study instructions and had recorded their voice while reading out the predefined sentences accordingly beyond the data-driven quality checks we had applied. Further, our approach did not allow to control for records' background noise (e.g., when participants were outside next to a road) or how they held their smartphone during the voice record. Since checking single raw audio files manually would be out of scope, future research could investigate additional approaches to check speech data quality directly on the device. Finally, in future work, smartphones could be used to log and immediately pre-process participants' everyday speech by using pre-trained language models to extract features of what they talk about (e.g., specific topics) directly on the device, too. Thereby, there would be no raw data transferred to a server, but potentially valuable information of language content could be also used for affect recognition.

1.7 Conclusion

In this work, we investigated if machine learning algorithms can recognize subjective affect experience from speech samples collected in the wild. Our cross-validated models that had been trained on extracted acoustic voice parameters and state-of-the-art word embeddings predicted affect experience on the dimension of overall arousal and contentedness, but not overall valence and sadness significantly better than chance. Overall, predictions were better when participants could talk freely (vs. reading out loud predefined emotional sentences). Also, speech content showed superior prediction performance compared to voice acoustics. Further, experimental and empirical findings suggest that emotional speech content did not affect predictions from voice acoustics (i.e., what someone talks about does not affect how well emotions can be predicted from voice cues alone). We discussed implications for theory and practice and further research opportunities. Finally, we discussed implications for the algorithmic monitoring of affect experience and raise issues regarding the protection of user privacy rights.

1.8 Author Contribution

In addition to myself, Ramona Schoedel (R.S.), Markus Bühner (M.B.), Clemens Stachl (C.S.), Florian Bemmann (F.B.), Gabriella Harari (G.H.), and Zachariah Marrero (Z.M.) contributed to this study. R.S., M.B., and G.H. acted as supervisors. F.B. created the logging software to collect the data for study 1.1. R.S. managed data collection of study 1.1. Z.M. assisted with preprocessing the raw data for study 1.2.

1.9 Acknowledgements

We thank Peter Ehrich and Dominik Heinrich for their support with the technical implementation of the on-device acoustic feature extraction for study 1. We thank ZPID for the support with data collection for study 1 and Sumer Vaid for the technical support in setting up the computational analysis environment for study 2.

1.10 Appendix

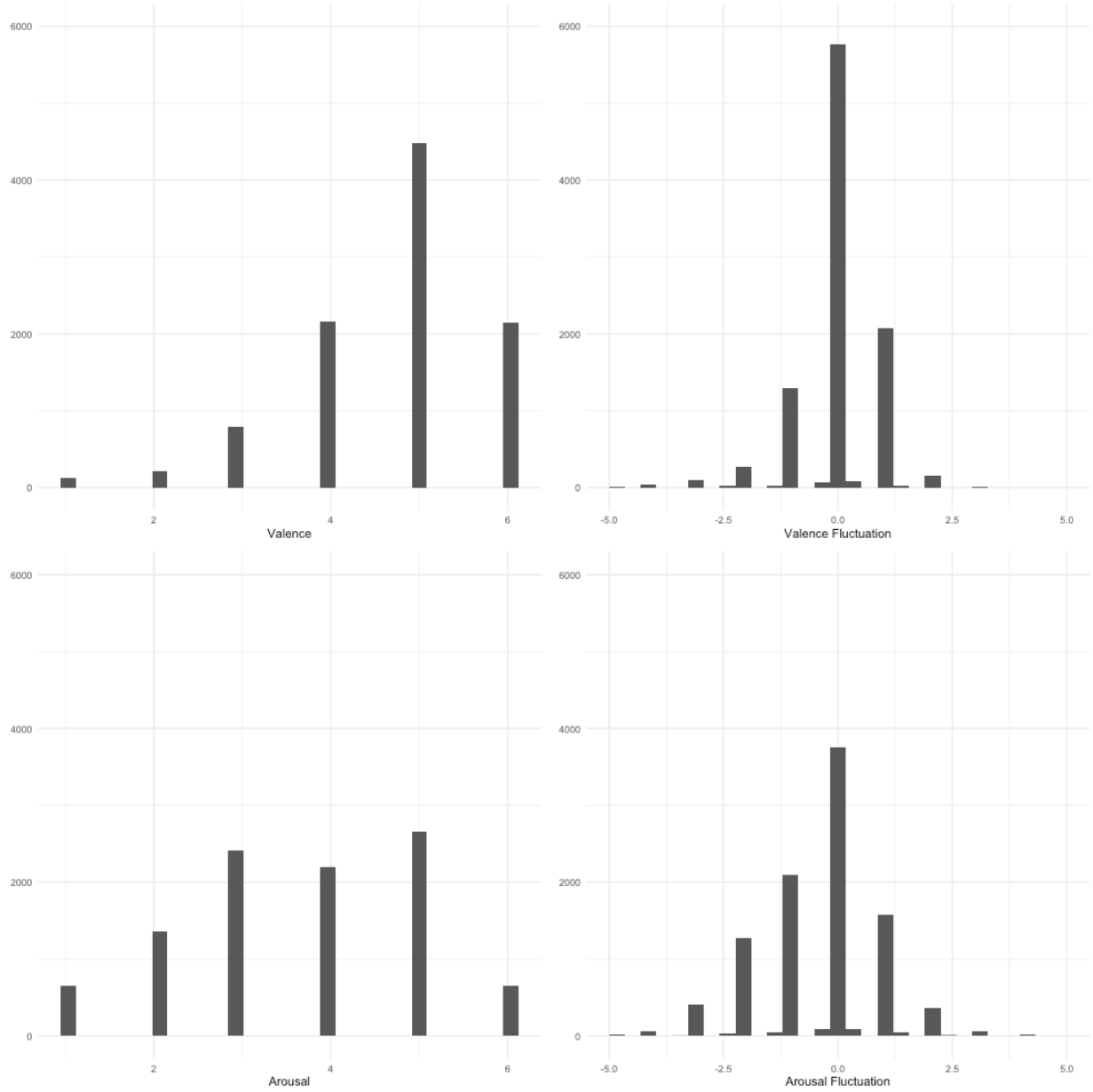


Figure 7: Distribution of affect measures in German data set (study 1.1).

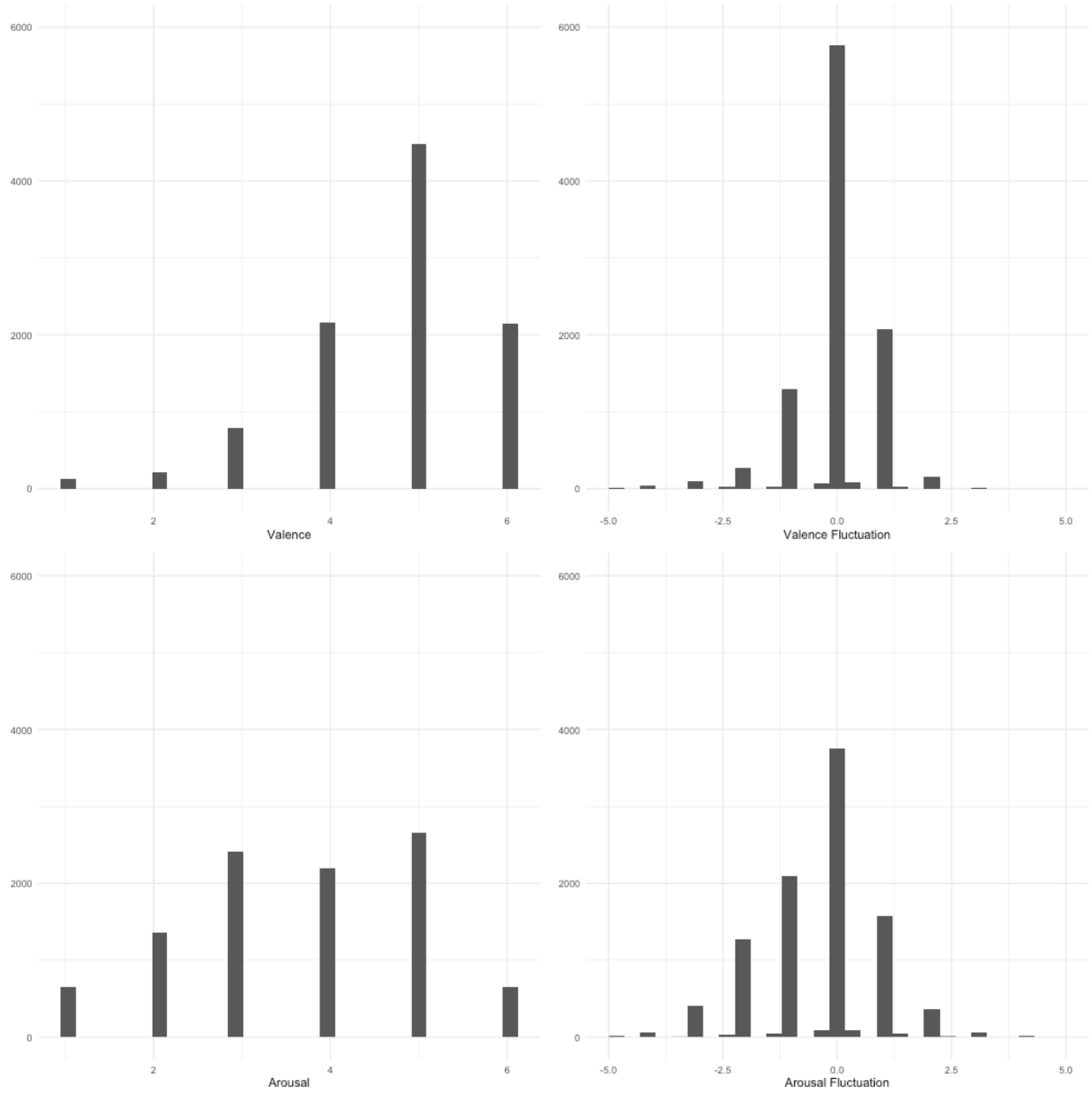
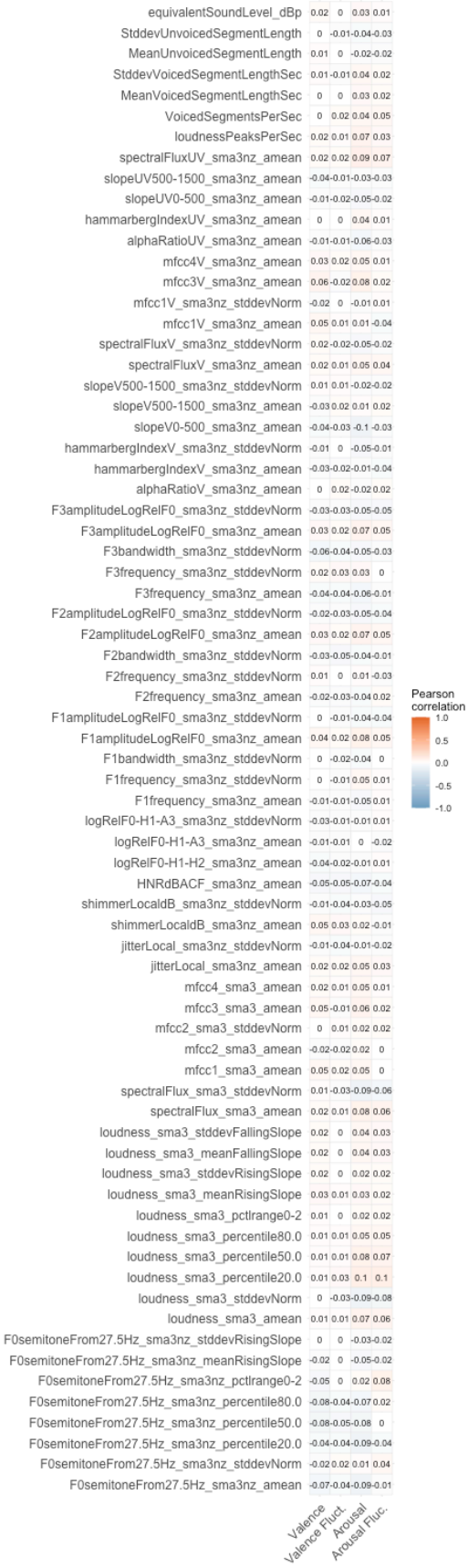


Figure 8: Distribution of affect measures in US data set (study 1.2).

\begin{figure}[H]

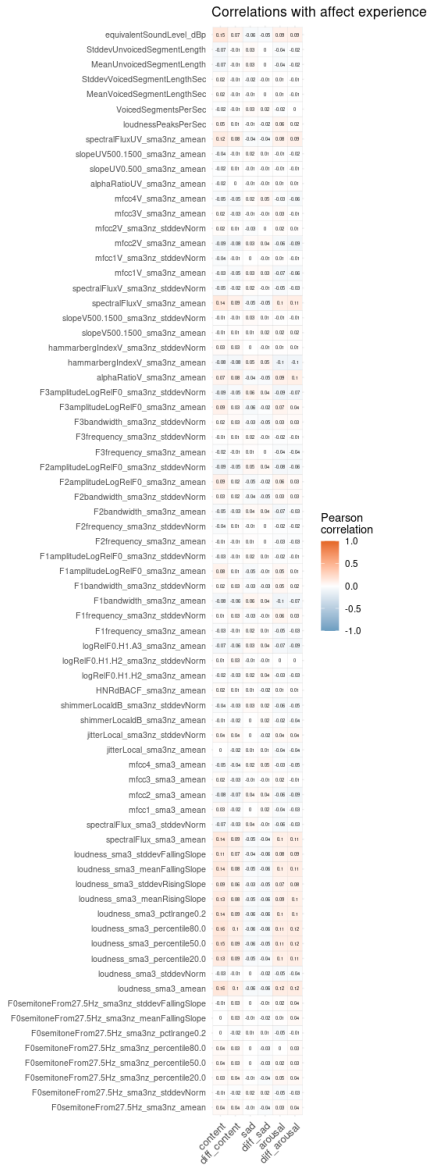

equivalentSoundLevel_dBp 0.02 0 0.03 0.01
StddevUnvoicedSegmentLength 0 -0.01 -0.04 -0.03
MeanUnvoicedSegmentLength 0.01 0 -0.02 -0.02
StddevVoicedSegmentLengthSec 0.01 -0.01 0.04 0.02
MeanVoicedSegmentLengthSec 0 0 0.03 0.02
VoicedSegmentsPerSec 0 0.02 0.04 0.05
loudnessPeaksPerSec 0.02 0.01 0.07 0.03
spectralFluxUV_sma3nz_amean 0.02 0.02 0.09 0.07
slopeUV500-1500_sma3nz_amean -0.04 -0.01 -0.03 -0.03
slopeUV0-500_sma3nz_amean -0.01 -0.02 -0.05 -0.02
hammarbergIndexUV_sma3nz_amean 0 0 0.04 0.01
alphaRatioUV_sma3nz_amean -0.01 -0.01 -0.06 -0.03
mfcc4V_sma3nz_amean 0.03 0.02 0.05 0.01
mfcc3V_sma3nz_amean 0.06 -0.02 0.08 0.02
mfcc1V_sma3nz_stddevNorm -0.02 0 -0.01 0.01
mfcc1V_sma3nz_amean 0.05 0.01 0.01 -0.04
spectralFluxV_sma3nz_stddevNorm 0.02 -0.02 -0.05 -0.02
spectralFluxV_sma3nz_amean 0.02 0.01 0.05 0.04
slopeV500-1500_sma3nz_stddevNorm 0.01 0.01 -0.02 -0.02
slopeV500-1500_sma3nz_amean -0.03 0.02 0.01 0.02
slopeV0-500_sma3nz_amean -0.04 -0.03 -0.1 -0.03
hammarbergIndexV_sma3nz_stddevNorm -0.01 0 -0.05 -0.01
hammarbergIndexV_sma3nz_amean -0.03 -0.02 -0.01 -0.04
alphaRatioV_sma3nz_amean 0 0.02 -0.02 0.02
F3amplitudeLogRelF0_sma3nz_stddevNorm -0.03 -0.03 -0.05 -0.05
F3amplitudeLogRelF0_sma3nz_amean 0.03 0.02 0.07 0.05
F3bandwidth_sma3nz_stddevNorm -0.06 -0.04 -0.05 -0.03
F3frequency_sma3nz_stddevNorm 0.02 0.03 0.03 0
F3frequency_sma3nz_amean -0.04 -0.04 -0.06 -0.01
F2amplitudeLogRelF0_sma3nz_stddevNorm -0.02 -0.03 -0.05 -0.04
F2amplitudeLogRelF0_sma3nz_amean 0.03 0.02 0.07 0.05
F2bandwidth_sma3nz_stddevNorm -0.03 -0.05 -0.04 -0.01
F2frequency_sma3nz_stddevNorm 0.01 0 0.01 -0.03
F2frequency_sma3nz_amean -0.02 -0.03 -0.04 0.02
F1amplitudeLogRelF0_sma3nz_stddevNorm 0 -0.01 -0.04 -0.04
F1amplitudeLogRelF0_sma3nz_amean 0.04 0.02 0.08 0.05
F1bandwidth_sma3nz_stddevNorm 0 -0.02 -0.04 0
F1frequency_sma3nz_stddevNorm 0 -0.01 0.05 0.01
F1frequency_sma3nz_amean -0.01 -0.01 -0.05 0.01
logRelF0-H1-A3_sma3nz_stddevNorm -0.03 -0.01 -0.01 0.01
logRelF0-H1-A3_sma3nz_amean -0.01 -0.01 0 -0.02
logRelF0-H1-H2_sma3nz_amean -0.04 -0.02 -0.01 0.01
HNRdBACF_sma3nz_amean -0.05 -0.05 -0.07 -0.04
shimmerLocaldB_sma3nz_stddevNorm -0.01 -0.04 -0.03 -0.05
shimmerLocaldB_sma3nz_amean 0.05 0.03 0.02 -0.01
jitterLocal_sma3nz_stddevNorm -0.01 -0.04 -0.01 -0.02
jitterLocal_sma3nz_amean 0.02 0.02 0.05 0.03
mfcc4_sma3_amean 0.02 0.01 0.05 0.01
mfcc3_sma3_amean 0.05 -0.01 0.06 0.02
mfcc2_sma3_stddevNorm 0 0.01 0.02 0.02
mfcc2_sma3_amean -0.02 -0.02 0.02 0
mfcc1_sma3_amean 0.05 0.02 0.05 0
spectralFlux_sma3_stddevNorm 0.01 -0.03 -0.09 -0.06
spectralFlux_sma3_amean 0.02 0.01 0.08 0.06
loudness_sma3_stddevFallingSlope 0.02 0 0.04 0.03
loudness_sma3_meanFallingSlope 0.02 0 0.04 0.03
loudness_sma3_stddevRisingSlope 0.02 0 0.02 0.02
loudness_sma3_meanRisingSlope 0.03 0.01 0.03 0.02
loudness_sma3_pctlrange0-2 0.01 0 0.02 0.02
loudness_sma3_percentile80.0 0.01 0.01 0.05 0.05
loudness_sma3_percentile50.0 0.01 0.01 0.08 0.07
loudness_sma3_percentile20.0 0.01 0.03 0.1 0.1
loudness_sma3_stddevNorm 0 -0.03 -0.09 -0.08
loudness_sma3_amean 0.01 0.01 0.07 0.06
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope 0 0 -0.03 -0.02
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope -0.02 0 -0.05 -0.02
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 -0.05 0 0.02 0.08
F0semitoneFrom27.5Hz_sma3nz_percentile80.0 -0.08 -0.04 -0.07 0.02
F0semitoneFrom27.5Hz_sma3nz_percentile50.0 -0.08 -0.05 -0.08 0
F0semitoneFrom27.5Hz_sma3nz_percentile20.0 -0.04 -0.04 -0.09 -0.04
F0semitoneFrom27.5Hz_sma3nz_stddevNorm -0.02 0.02 0.01 0.04
F0semitoneFrom27.5Hz_sma3nz_amean -0.07 -0.04 -0.09 -0.01
Valence
Valence Fluct.
Arousal
Arousal Fluc.
Pearson correlation
1.0
0.5
0.0
-0.5
-1.0

{

}

\caption{Pearson correlations of voice features from the eGeMAPS voice feature set with between-person differences and within-person fluctuations in momentary self-reported affect experience on the dimensions of valence and arousal. Those voice features are displayed for which the 95% confidence interval of the correlation coefficient does not contain zero for any of the outcomes. This figure is based on data from study 1.1.} \end{figure}

\begin{figure}[H]



{

}

\caption{Pearson correlations of voice features from the eGeMAPS voice feature set with between-person differences and within-person fluctuations in momentary self-reported affect experience on the dimensions of contentedness, sadness, and arousal. Those voice features are displayed for which the 95% confidence interval of the correlation coefficient does not contain zero for any of the outcomes. This figure is based on data from study 1.2.} \end{figure}

2 References

Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., … Amir, N. (2011). The Automatic Recognition of Emotions in Speech. In *Cognitive Technologies* (pp. 71–99). http://doi.org/10.1007/978-3-642-15184-2_6

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion (Washington, D.C.)*, *12*(5), 1161–1179. http://doi.org/10.1037/a0025827

Ben-David Boaz M., Multani Namita, Shakuf Vered, Rudzicz Frank, & van Lieshout Pascal H. H. M. (2016). Prosody and Semantics Are Separate but Not Separable Channels in the Perception of Emotional Speech: Test for Rating of Emotions in Speech. *Journal of Speech, Language, and Hearing Research*, *59*(1), 72–89. http://doi.org/10.1044/2015_JSLHR-H-14-0323

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, *19*(84), 1–5.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Interspeech* (Vol. 5, pp. 1517–1520).

Carlier, C., Niemeijer, K., Mestdagh, M., Bauwens, M., Vanbrabant, P., Geurts, L., … Kuppens, P. (2022). In Search of State and Trait Emotion Markers in Mobile-Sensed Language: Field Study. *JMIR Mental Health*, *9*(2), e31724. http://doi.org/10.2196/31724

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. http://doi.org/10.21437/SpeechProsody.2018-145

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., … Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. http://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia - MM '10* (p. 1459). Firenze, Italy: ACM Press. http://doi.org/10.1145/1873951.1874246

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *2008 IEEE International Conference on Multimedia and Expo* (pp. 865–868). http://doi.org/10.1109/ICME.2008.4607572

Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, *121*, 364–374. http://doi.org/10.1016/j.jbusres.2020.09.020

Huang, Z., & Epps, J. (2018). Prediction of Emotion Change From Speech. *Frontiers in ICT*, *5*.

Kjell, O., Giorgi, S., & Schwartz, H. A. (2021, April). Text: An R-package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning. PsyArXiv. http://doi.org/10.31234/osf.io/293kt

Koch, T., & Schoedel, R. (2021). Predicting Affective States from Acoustic Voice Cues Collected with Smartphones. http://doi.org/10.23668/psycharchives.4454

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., … Bischl, B. (2019). Mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. http://doi.org/10.21105/joss.01903

Lin Yi, Ding Hongwei, & Zhang Yang. (2020). Prosody Dominates Over Semantics in Emotion Word Processing: Evidence From Cross-Channel and Cross-Modal Stroop Effects. *Journal of Speech, Language, and Hearing Research*, *63*(3), 896–912. http://doi.org/10.1044/2020_JSLHR-19-00258

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. http://doi.org/10.48550/arXiv.1907.11692

Mandell, J. (2020). Spotify Patents A Voice Assistant That Can Read Your Emotions. *Forbes*. https://www.forbes.com/sites/joshmandell/2020/03/12/spotify-patents-a-voice-assistant–that-can-read-your-emotions/.

Marrero, Z. N. K., Gosling, S. D., Pennebaker, J. W., & Harari, G. M. (2022). Evaluating voice samples as a potential source of information about personality. *Acta Psychologica*, *230*, 103740. http://doi.org/10.1016/j.actpsy.2022.103740

Matero, M., Hung, A., & Schwartz, H. A. (2022). Evaluating Contextual Embeddings and their Extraction Layers for Depression Assessment. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (pp. 89–94). Dublin, Ireland: Association for Computational Linguistics. http://doi.org/10.18653/v1/2022.wassa-1.9

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A Method for the Naturalistic Observation of Daily Social Behavior. *Current Directions in Psychological Science*, *26*(2), 184–190. http://doi.org/10.1177/0963721416680611

Miller, G. (2012). The Smartphone Psychology Manifesto. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(3), 221–237. http://doi.org/10.1177/1745691612441215

Milling, M., Pokorny, F., Bartl-Pokorny, K., & Schuller, B. (2022). Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell. *Frontiers in Digital Health*, *4*, 886615. http://doi.org/10.3389/fdgth.2022.886615

Molnar, C. (2019). *Interpretable Machine Learning*.

Müller, S. R., Chen, X. L., Peters, H., Chaintreau, A., & Matz, S. C. (2021). Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports*, *11*(1), 14007. http://doi.org/10.1038/s41598-021-93087-x

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256. http://doi.org/10.1016/S0167-6393(02)00084-5

Schoedel, R., & Oldemeier, M. (2020). Basic Protocol: Smartphone Sensing Panel Study. http://doi.org/10.23668/psycharchives.2901

Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99. http://doi.org/10.1145/3129340

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., … Evanini, K. (2016). *The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language* (p. 2005). http://doi.org/10.21437/Interspeech.2016-129

Schwartz, R., & Pell, M. D. (2012). Emotional Speech Processing at the Intersection of Prosody and Semantics. *PLoS ONE*, *7*(10), e47279. http://doi.org/10.1371/journal.pone.0047279

Sridhar, K., & Busso, C. (2022). Unsupervised Personalization of an Emotion Recognition System: The Unique Properties of the Externalization of Valence in Speech. *IEEE Transactions on Affective Computing*, *13*(4), 1959–1972. http://doi.org/10.1109/TAFFC.2022.3187336

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364–387. http://doi.org/10.1037/pspp0000244

Vlahos, J. (2019). *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think*. Eamon Dolan Books.

Vogt, T., André, E., & Wagner, J. (2008). Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation. In C. Peter & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction* (Vol. 4868, pp. 75–91). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-85099-1_7

Weidman, A. C., Sun, J., Vazire, S., Quoidbach, J., Ungar, L. H., & Dunn, E. W. (2020). (Not) hearing happiness: Predicting fluctuations in happy mood from acoustic cues using machine learning. *Emotion (Washington, D.C.)*, *20*(4), 642–658. http://doi.org/10.1037/emo0000571

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00292

Wilting, J., Krahmer, E. J., & Swerts, M. G. J. (2006). Real vs. Acted emotional speech. *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)*.

Wright, M. N., & Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. http://doi.org/10.18637/jss.v077.i01

Wu, C., Fritz, H., Bastami, S., Maestre, J. P., Thomaz, E., Julien, C., … Nagy, Z. (2021). Multi-modal data collection for measuring health, behavior, and living environment of large-scale participant cohorts. *GigaScience*, *10*(6). http://doi.org/10.1093/gigascience/giab044