Semantic content outperforms speech prosody in predicting affective experience in naturalistic settings

Abstract

Many commercial products use algorithms to recognize affect based on speech prosody (i.e., voice acoustics). These algorithms are typically trained on enacted or labeled speech samples collected in lab settings. However, they are used to infer affective experiences occurring in everyday life. Here, we investigate whether the experience of affective states can be predicted from speech samples collected using smartphones in naturalistic settings. In two field studies (experimental Study 1: $N = 409$; observational Study 2: $N = 687$), we collected 25,403 speech samples from participants along with their self-reported affective experiences. Machine learning analyses show that prosody reveals limited affective information ($r_{\text{md}} = .17$) and is outperformed by semantic content ($r_{\text{md}} = .33$). Our findings demonstrate the importance of semantic content and challenge whether previously reported prediction performances for affective expression from prosody in controlled settings generalize to the recognition of subjective affective experience in naturalistic settings.

Keywords: Affect, Voice, Speech, Prosody, Machine Learning

Word count: 5144

Semantic content outperforms speech prosody in predicting affective experience in naturalistic settings

The prosody of a voice, such as acoustic tone, pitch, and rhythm, serves as a primary means for conveying emotional information (Kraus, 2017; Scherer, 2003). Decades of research investigated how speech prosody varies with affect (Larrouy-Maestri, Poeppel, & Pell, 2024), finding, for example, that higher pitch and faster speech rates are associated with excitement or stress (Banse & Scherer, 1996). More recently, algorithms trained to automatically recognize affect and affective disorders from speech prosody offer promising potential applications in the domains of health care, human-machine interaction, education, and business (Hildebrand et al., 2020; Milling, Pokorny, Bartl-Pokorny, & Schuller, 2022; Vlahos, 2019). The widespread adoption of voice assistants like Amazon's Alexa and Apple's Siri has also fueled commercial interest in developing algorithms that can detect affect in everyday life. These algorithms aim to quantify at scale how patients, customers, and employees feel in a given moment, often to provide personalized recommendations, feedback, and services (e.g., Matz & Netzer, 2017; Seiferth et al., 2023).

Algorithmic recognition of affective expression in the lab versus experience in the wild

Existing affect recognition algorithms are predominantly trained on enacted emotional speech (from professional actors in lab settings, Bänziger, Mortillaro, & Scherer, 2012; Schuller, 2018; Vogt, André, & Wagner, 2008) or on labeled speech samples (such as TV clips, Grimm, Kroschel, & Narayanan, 2008). Using such training data generally results in high performance for the algorithmic recognition of, for example, affective arousal ($r_{\max} = .81$) and valence ($r_{\max} = .68$) in controlled settings (Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013). However, these data, which are used to train the algorithms, differ in two important respects from the data that are actually processed by the algorithms when they are applied in the real world.

First, the data used to train such algorithms consists of actors' portrayals of affective states or raters' labels of existing voice samples that both rely on folk theories of how affect is expressed; for instance, how does a sad person sound when they talk? (Batliner et al., 2011; Schuller, 2018; Wilting, Krahmer, & Swerts, 2006). However, those prototypical affective expressions are not necessarily reflective of people's subjective affective *experience*; that is, their genuine subjective affective state. For example, someone might feel nervous about giving a presentation, but choose to express themselves by talking calmly and confidently. Investigating affective expression is invaluable for understanding emotional *communication* (Ekman & Friesen, 1969), but the main purpose of such algorithms is to detect people's subjective affective *experience*. This is important for both researchers and commercial entities who want to gain a deeper understanding of how individuals, such as patients or customers, genuinely feel, as opposed to merely observing how they outwardly express themselves.

The second way the data used to train current algorithms differ from instances in which they will be used is whether speech data has been produced and collected in laboratory or real-world settings. In controlled lab settings, actors enact predefined affective states or desired emotions are elicited in participants and their speech is recorded, allowing for the investigation of affective speech in standardized conditions using controlled stimuli (e.g., Cowen, Laukka, Elfenbein, Liu, & Keltner, 2019). However, speech data collected in lab settings limit ecological validity. Algorithms trained on such data may perform poorly in real-world situations, where conditions differ significantly from controlled environments. For example, lab recordings are made in isolated rooms without background noise, unlike the complex auditory environments of everyday life (e.g., noisy cafés or restaurants). As a result, these algorithms might fail to accurately recognize affect in naturalistic settings, leading to unreliable results and misinterpretations in practical applications.

The focus on affective expression (versus affective experience) and on lab-based (versus real world) speech samples that characterized past research raises concerns about the generalizability of the performance of existing algorithms. Specifically, the speech data on which these algorithms are being trained do not match the speech data encountered in the settings in which they are deployed. This raises the question: Are algorithms capable of accurately detecting subjective affective experience in everyday life, and with comparable degrees of accuracy as they have shown for affective expression in the lab?

Prosody versus semantics for affect recognition from speech

In addition to prosody, the semantic content (i.e., meaning of the words) of speech plays a central role in conveying affective information. Prosody and semantics interact dynamically, working together to transmit affective cues through speech (Ben-David, Multani, Shakuf, Rudzicz, & van Lieshout, 2016). Prior research suggests that prosody often takes precedence over semantics in how humans perceive affective states (Ben-David, Multani, Shakuf, Rudzicz, & van Lieshout, 2016; Lin Yi, Ding Hongwei, & Zhang Yang, 2020). In the same fashion, prior work indicates that algorithms might also prioritize prosodic cues over semantics for the recognition of affect from speech (El Ayadi, Kamel, & Karray, 2011; Polzehl, Schmitt, Metze, & Wagner, 2011; Schuller, Rigoll, & Lang, 2004). However, as noted above, these findings are based on algorithms trained to recognize affect expression (not experience) in controlled lab settings. Consequently, what remains unknown is whether algorithms rely more on prosodic features or semantic content when recognizing affective experience from spontaneous speech in real-world settings.

In the present work, we address the existing gaps in algorithmic affect recognition from speech, namely (1) the generalization of algorithms' prediction performance from affective expression in the lab to affective experience in the real world, and (2) the predictive power of prosody versus semantics for affect inferences from speech. To investigate the algorithmic recognition of affective experience in everyday life, we conducted two large-scale field studies in two different countries: a field experiment with German-speaking participants in Germany

(Study 1) and an observational field study with English-speaking participants in the United States (Study 2). In both studies, we used smartphone applications to collect experience sampling reports and speech samples via the built-in microphone as people went about their daily lives. Using this approach, we collected a total of 25,403 speech samples, paired with self-reported affective experience from 1,096 participants.

The first study was a field experiment. We manipulated what people said in speech samples collected in naturalistic settings to conservatively test the predictive power of voice prosody for algorithmic recognition of affective experience, while controlling for semantic content. Participants were instructed to read out pre-tested scripted sentences with varying emotional sentiment (i.e., positive, negative, and neutral) while recording their voice. From the extracted prosodic features we then predicted self-reported affective states on the dimensions of arousal and valence using machine learning. In the second study, we collected spontaneous speech samples in naturalistic settings to investigate the predictive power of voice prosody and semantic content for algorithmic recognition of affective experience. Participants were prompted to talk spontaneously about their current situation, thoughts, and feelings while recording their voice. Just as in Study 1, we subsequently predicted self-reported affective states on the dimensions of arousal as well as contentment and sadness (for emotional valence) from voice prosody to investigate the predictive power of prosody in spontaneous speech. Additionally, we used the semantic content captured by word embeddings from those speech samples to predict momentary affective states and compare the models' performance to those trained solely on prosody.

Results

In this section, we report on the performance of machine learning models predicting self-reported affective states from prosodic and semantic features derived from scripted and spontaneous speech samples collected via smartphones. While Figure 1 provides an overview of models' performance across all predictions and algorithms, we report on the average performance of the best performing algorithm in the text.

Prosody reveals limited affective information in scripted and spontaneous speech

When participants read out scripted content (Study 1), the prediction of momentary affective arousal from voice prosody was low on average ($r_{\text{md}} = 0.17$, $r_{\text{sd}} = 0.08$), yet significantly better than chance levels. The prediction of momentary affective valence from voice prosody was on average even lower than for arousal and was not significantly better than chance ($r_{\text{md}} = 0.03$, $r_{\text{sd}} = 0.07$). We analyzed whether the experimentally altered sentiment (positive/ neutral/ negative) of the scripted sentences had an effect on affect predictions from prosody. We found no significant differences in prediction errors across the three sentence conditions for arousal (F(2,49702) = 0.01, p = .500) and valence (F(2,49702) = 0.70, p = .993) predictions suggesting that sentences' sentiment did not influence affect predictions from prosodic features.

Models trained on prosodic features from participants talking spontaneously about their current situation, thoughts, and feelings (Study 2) also yielded a low average prediction performance overall. Still, the prediction of momentary arousal ($r_{\text{md}} = 0.12$, $r_{\text{sd}} = 0.04$) and contentment ($r_{\text{md}} = 0.15$, $r_{\text{sd}} = 0.05$) were significantly better than chance. Predictions of sadness were not better than the baseline ($r_{\text{md}} = 0.03$, $r_{\text{sd}} = 0.05$). To provide insights into which prosodic features could be associated with affective states, we include an overview of the importance of prosodic features in Elastic Net models and report on the correlations of voice features with momentary affective experience in the repository.

Semantic content outperforms prosody in predicting subjective affective experience

Machine learning models trained on semantic content in the form of word embeddings extracted from participants' spontaneous speech (Study 2) yielded good performance results overall that were significantly better than chance, outperforming models trained on voice prosody. Specifically, we found semantic content to be, on average, more predictive of the momentary experience of arousal ($r_{\mathrm{md}} = 0.31$, $r_{\mathrm{sd}} = 0.04$), contentment ($r_{\mathrm{md}} = 0.33$, $r_{\mathrm{sd}} = 0.03$), and sadness ($r_{\mathrm{md}} = 0.22$, $r_{\mathrm{sd}} = 0.03$) than voice prosody. Models trained on both prosodic features and word embeddings showed a similar prediction performance to those trained on word embeddings alone, suggesting that the predictions were primarily driven by semantic information.

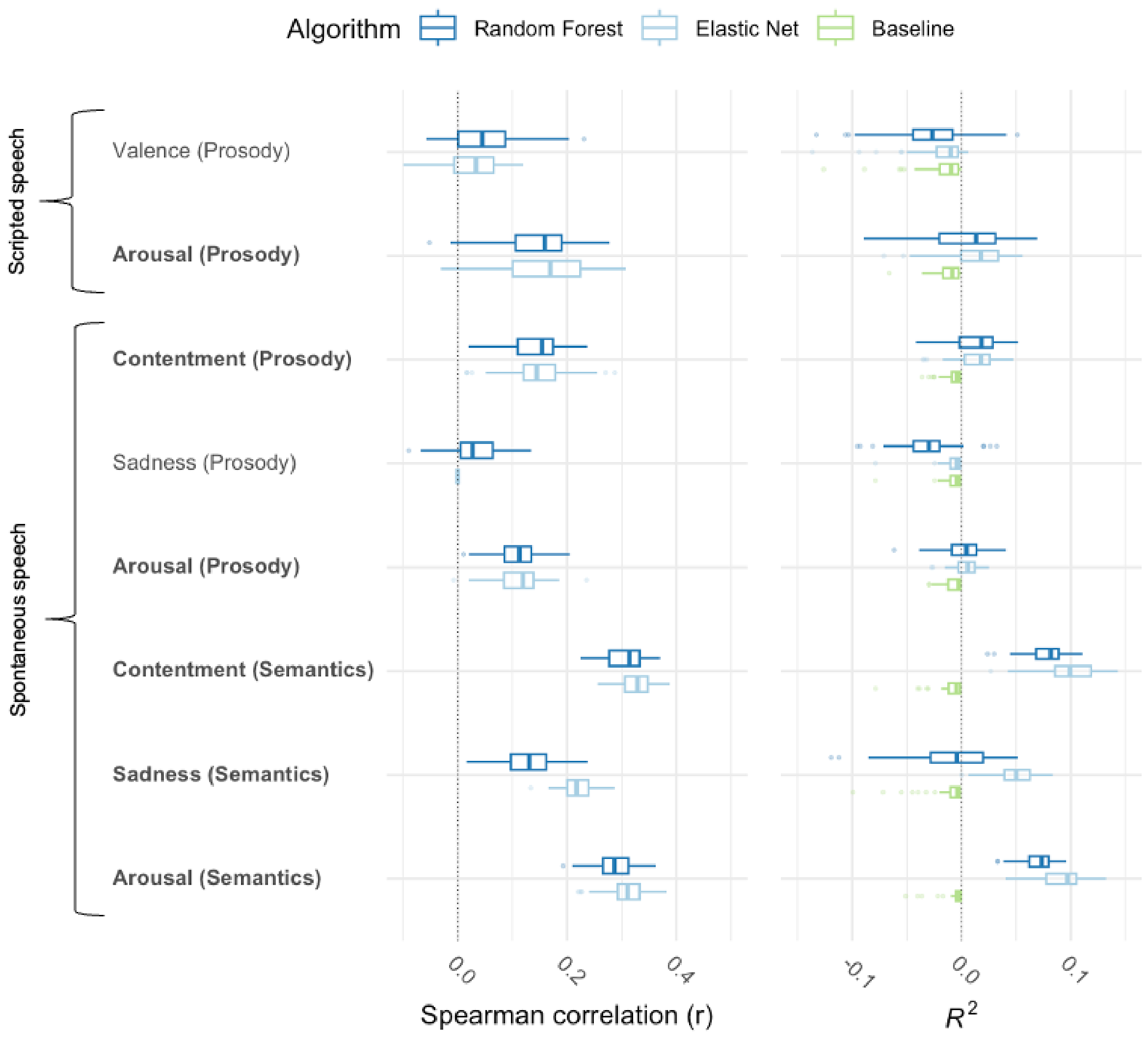

Figure 1. Box and whisker plot of out-of-sample prediction performance measures from five times repeated ten-fold cross-validation for each affect measure. Feature sets (either prosodic or semantic features) for model training are shown in parentheses. Names of models that are significantly better than chance are displayed in boldface. Pearson correlation is not available for baseline models because it predicts a constant value, for which correlation measures are not defined.

Discussion

Affect-recognition algorithms developed in research and commercial tools are mainly trained on prosody extracted from enacted or labeled speech samples collected in controlled lab settings. Yet, these algorithms are deployed to detect people's subjective affective experience in naturalistic settings. In the present work, we investigated whether subjective momentary affective states are predictable from prosodic and semantic features in real-world speech samples collected with smartphones. In contrast to prior work, which was based on affective expressions from enacted speech data from the lab, our findings suggest that prosody reveals only limited information about affective experience and is outperformed by semantic content in the real world.

The first major finding to emerge from this work is that prosody provided limited information about affective experience in naturalistic settings. Machine learning models trained on prosodic features achieved a lower prediction performance for affective experience (up to $r_{\text{md}} = .17$) compared to prior work on the algorithmic prediction of affective expression in controlled lab settings (up to $r_{\text{max}} = .81$, e.g., Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013), which might have also been susceptible to overfitting to the training data (Aishwarya, Kaur, & Seemakurthy, 2024). This result supports the notion that recognizing real-life affective experience from voice cues algorithmically is inherently more challenging than detecting affective expression in the lab (Vogt, André, & Wagner, 2008). Nevertheless, past research predicting affective experience from speech in the wild reported similar prediction performance (Carlier et al., 2022; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). Moreover, the prediction of arousal from prosody was successful across prediction models whereas the prediction of valence yielded significant predictions only for contentment, but not for overall valence and sadness. This finding is in line with prior work showing that valence is more challenging to infer algorithmically than is arousal due to its subjective nature (Sridhar & Busso, 2022). Thus, our findings raise questions about the generalizability of the high prediction performance from past research (e.g., Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013), which focus on recognizing affective expression from speech data collected in lab settings. Specifically, it seems that the performance obtained from lab-based studies may not extend to the recognition of affective experience from speech prosody in everyday life.

Another reason why prediction performance is lower in real-world data sets could be that they contain fewer instances of extreme affective experience than those used in lab studies focusing on enacted or labeled speech. Consequently, our predictions targeted "normal" everyday affective states with only a few cases of extreme affect experience, which are inherently less intense than the short-lived emotions that have been investigated in prior work (Scherer, 2003), and thus present a greater challenge for algorithmic recognition (Vogt, André, & Wagner, 2008).

To test whether a broader set of prosodic features would improve prediction performance, we reran predictions using the 2016 Interspeech Computational Paralinguistic Challenge set of 6,737 prosodic features (versus 88 features in the eGeMAPS set used in the main analyses) (Eyben et al., 2016; Schuller et al., 2016). In line with prior research, the larger voice feature set did not yield better predictions of momentary affective experience (Weidman et al., 2020), suggesting that the limited predictive power of prosody was not due to the parsimonious size of the prosodic feature set (see repository for details).

The second major finding to emerge from this work is that semantic content captured by word embeddings outperformed prosody in predicting affective experience in spontaneous speech. This insight suggests that speech content may be more informative than prosodic cues for algorithms recognizing experienced affect in everyday life. It contrasts prior research suggesting that speech prosody is more relevant than semantics for affect inferences by humans (Ben-David, Multani, Shakuf, Rudzicz, & van Lieshout, 2016; Lin Yi, Ding Hongwei, & Zhang Yang, 2020) and algorithms trained on lab data (El Ayadi, Kamel, & Karray, 2011;

Polzehl, Schmitt, Metze, & Wagner, 2011; Schuller, Rigoll, & Lang, 2004). However, our results align with primary research that found semantic content to be more predictive than prosodic cues in algorithmically predicting momentary subjective affective experience (Carlier et al., 2022; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020). As an explanation, we argue that algorithms are potentially better at detecting signals from the structured nature of text, in contrast to the complexity of prosody, such as subtle nuances of intonation (e.g., by stressing a single word). Nonetheless, participants in our study were prompted to speak spontaneously about their current situations, thoughts, and feelings, which can be expected to reveal some affective information. Further research is needed to examine how the themes people talk about (e.g., explicitly describing how one feels, which is likely to carry a rich affective signal, versus describing the physical environment, which may contain minimal affective content) impact prediction performance.

Finally, we replicated analyses from past work that investigated predictions of fluctuations in affective states from speech prosody and semantic content. The results were consistent with prior work, indicating that the prediction performance for within-person fluctuations in affective states is lower than the performance of models predicting between-person differences (Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020).

The findings of this work are limited in three ways. First, there are some differences between the two studies that might affect the comparability of results: In particular, the studies differed in terms of the scales used to assess affective experience (see methods section for details), the sampling strategy (representative quota sample vs. undergraduate cohort), language and culture (Germany vs. USA), and smartphone platforms (Android only vs. Android and iOS). We sought to minimize the effects of these differences by, where possible, using consistent methodologies and statistical approaches across studies. Nonetheless, the findings might be considered transferable but not directly comparable across studies.

Second, both data sets were collected in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) countries (Henrich, Heine, & Norenzayan, 2010). Plenty of prior research has pointed to cultural differences in the experience and expression of affect (Lim, 2016), including cultural variations in prosodic affect markers in the voice (Brooks et al., 2023; Laukka & Elfenbein, 2021; van Rijn & Larrouy-Maestri, 2023). Therefore, the generalizability of the current work should be considered as limited to this cultural area. Future studies should seek to broaden participant inclusion to diverse cultural contexts, especially non-Western countries (Phan, Modersitzki, Gloystein, & Müller, 2023).

Third, in contrast to prior work using passive speech sensing, for example via the Electronically Activated Recorder (EAR) (Mehl, 2017; Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020), participants had to actively log their speech via their smartphone in the present work. As a consequence, participants might not have spoken as naturally as they would have in a naturalistic conversation. Further, they might have made the audio records only in selected situations that were suitable for making an audio record, for example when they were alone in a quiet place.

Conclusion

In this work, we investigated whether machine learning algorithms can recognize subjective affective experience from naturalistic speech samples collected in everyday life via smartphones. Extracted prosodic voice parameters provided only limited affective information and were outperformed by semantic content as captured in word embeddings. Our findings challenge whether the optimistic prediction performance results from prior research on the recognition of affective expression (e.g., enacted and labeled speech) from lab settings generalize to the recognition of subjective affective experience in everyday speech. The ability to detect how people genuinely feel, as opposed to merely observing what they outwardly express, is crucial to gaining a deeper understanding of human emotions and guiding the development of technologies that are better placed to serve and respond to our emotional needs.

Methods

Data collection

The present work consists of two studies leveraging machine learning, each based on a separate dataset. The workflow of data collection, processing, and predictive modeling is described in detail in the following section and illustrated in Figure 2.

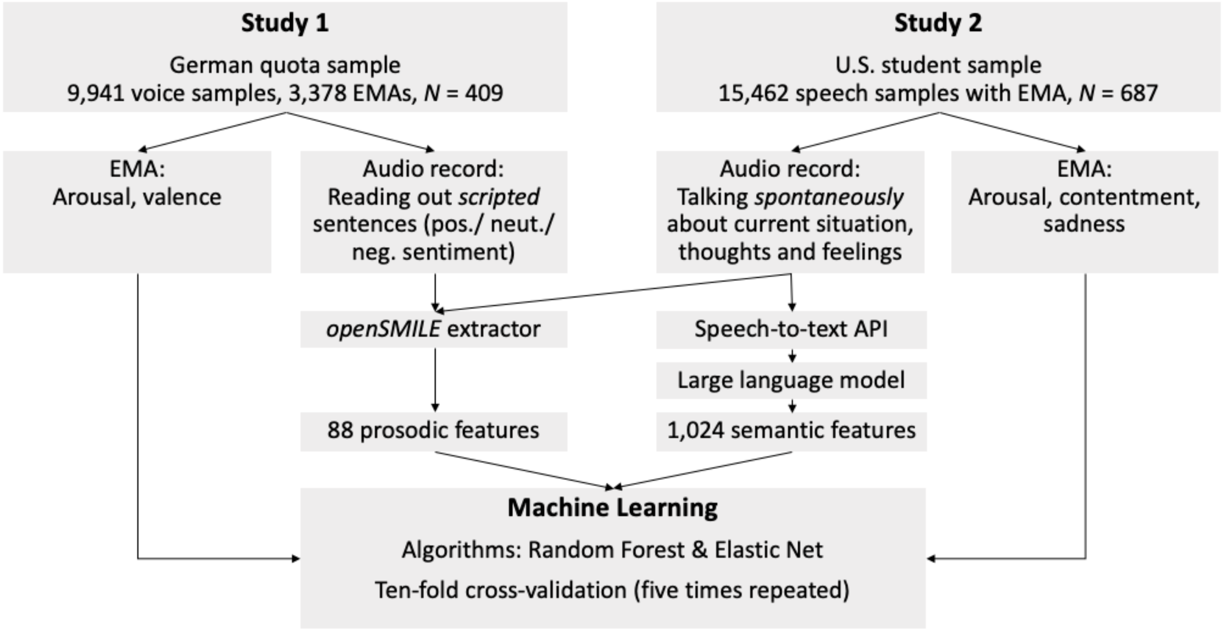
Study 1
German quota sample
9,941 voice samples, 3,378 EMAs, N = 409
Study 2
U.S. student sample
15,462 speech samples with EMA, N = 687
EMA: Arousal, valence
Audio record: Reading out scripted sentences (pos./ neut./ neg. sentiment)
Audio record: Talking spontaneously about current situation, thoughts and feelings
EMA: Arousal, contentment, sadness
openSMILE extractor
Speech-to-text API
Large language model
88 prosodic features
1,024 semantic features
Machine Learning
Algorithms: Random Forest & Elastic Net
Ten-fold cross-validation (five times repeated)

Figure 2. Flow diagram illustrating data collection, processing, and predictive modeling in Study 1 and Study 2.

Study 1. Data collection for Study 1 was part of a large panel study (from May 15 until November 14, 2020) using the *PhoneStudy* research app (Schoedel & Oldemeier, 2020). Data collection was approved by the ethics committee of the psychology department at a German university and all procedures adhered to the General Data Protection Regulation (GDPR). We recruited a quota sample of $N = 850$ participants representative of the German population regarding age, gender, education, income, confession, and relationship status. Participants had to be between 18 and 65 years old, fluent in German, and in the possession of an Android smartphone (Android 5 or higher) for private usage as a sole user. The study also 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 on their personal smartphone as part of an experience sampling procedure. Here, self-reported momentary affective valence and arousal were assessed in two separate items on six-point Likert scales: Valence ranging from "very unpleasant" (1) to "very pleasant" (6) and arousal ranging from "very inactive" (1) to "very activated" (6). The experience sampling procedure also covered other psychological properties (i.e., sleep quality, stress, situational characteristics) that had been used in other research (große Deters & Schoedel, 2024; Schoedel et al., 2023).

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 54 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."). In each measurement instance, participants were instructed to read out three positive, neutral, and negative sentences. The order of the categories was randomized per experience sampling instance. For each emotional content category, three sentences were randomly drawn (with replacement) from the respective sets of sentences in the database. The experimental manipulation of these emotional semantic categories allowed us to control for the content spoken by our participants and at the same time enabled 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 per condition, determined by reading the sentences extremely fast and extremely slow and recording the times. Once the audio record had been completed, we used the widely adopted OpenSMILE open-source algorithm (Eyben, Wöllmer, & Schuller, 2010) to automatically extract acoustic features directly on the participant's device. Here, we used the extended Geneva Minimalistic Acoustic Parameters Set (eGeMAPS) that comprises 88 acoustic features (Eyben, Wöllmer, & Schuller, 2010) and the 2016 Interspeech Computational Paralinguistic Challenge set of 6,737 acoustic features (Schuller et al., 2016). Those feature sets have been used in a range of prior studies on affect recognition from speech (van Rijn & Larrouy-Maestri, 2023; Weidman et al., 2020). We used the parsimonious eGeMAPS feature set for our main analyses. After feature extraction, the voice records were automatically deleted and only the extracted voice features were transmitted to our servers. As a result, there were three sets of voice features per experience sampling instance (one per sentiment condition).

We collected 11187 audio logs with respective affect self-reports from 577 participants. Participants made on average 19.39 (SD = 12.37) voice records. We excluded data from 158 participants with less than ten voice samples in total and the data from eight participants who had no variance in all their valence and arousal scores across all their observations. Finally, we excluded 163 voice logs from 22 participants because the respective voice indicators suggested that no human voice was recorded (mean voicing score < 0.5, voiced segments per second $= 0$, mean voiced segment length $= 0$). This left us with a final data set of 9941 voice samples with corresponding acoustic features from 3378 experience sampling instances for valence and arousal from 409 participants (49.48% female, $\text{Age}_{\text{M}} = 42.97$ years). In the final sample, there were on average 24.31 (SD = 10.54) voice logs with 8.26 (SD = 3.53) corresponding affect self-reports per participant. Overall self-reported valence was positive (M = 4.70, SD = 1.04) and overall arousal was slightly geared towards activation (M = 3.67, SD = 1.34). The distribution of valence and arousal ratings is provided in the repository.

Study 2. Data collection for Study 2 was part of the UT1000 Project at a public university 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) in introductory psychology received four short experience sampling questionnaires per day where they could also make speech records at the end. Here, self-reported arousal (assessed on a five-point Likert scale ranging from “low energy” (1) to “high energy” (5)), contentment, and sadness were assessed in separate items on four-point Likert scales (ranging from “not at all” (1) to “very much” (4)) among other psychological properties as part of an experience sampling procedure (Wu et al., 2021). In Study 2, we captured emotional valence on two items (contentment and sadness) instead of one as done in Study 1. According to the affect grid, contentment and sadness have a comparable low level of arousal and an opposing emotional valence (Posner, Russell, & Peterson, 2005; Russell, 1980).

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 responses to this prompt are analyzed in the present study. Any parts of the record that did not contain speech were cut out before further analysis since the focus of this work is on affect in human speech. The collected speech samples had also been used in another research project that describes the data collection procedure in more detail (Marrero, Gosling, Pennebaker, & Harari, 2022).

In total, we collected 23482 audio logs with corresponding affect self-reports from 980 participants. Participants made on average 23.96 ($SD = 18.40$) speech records with corresponding affect self-reports. We followed the identical procedure to filter the data as in Study 1: First, we excluded the data from 281 participants with less than ten audio records in total and from another participant who had no variance in all their self-reports across all their experience samples. To ensure comparability of the two studies with regard to the length of speech samples, we removed 6,871 speech transcripts that contained less than 15 words and were less than four seconds long which is equivalent to the length of the sentences that had been read out in Study 1. Acoustic features indicated that human voice had been recorded in all of the remaining speech samples.

This procedure left us with a final data set of 15462 speech samples with an average length of 52.37 words (*SD* = 25.49) and duration of 34.33 seconds (*SD* = 13.13) with corresponding experience-sampled self-reports on momentary affective experience from 687 participants (62.49% female, Age_{M} = 18.58 years). In the final sample, there were on average 22.51 (*SD* = 14.10) speech samples with corresponding affect self-reports per participant. Overall participants reported balanced experienced contentment (*M* = 1.68, *SD* = 0.86) and low sadness (*M* = 0.48, *SD* = 0.75). Overall arousal was balanced out (*M* = 1.97, *SD* = 0.95). The distribution of self-reported arousal, contentment, and sadness scores is provided in the repository.

Replicating the approach from Study 1, we extracted the extended Geneva Minimalistic Acoustic Parameters Set (eGeMAPS) and the 2016 Interspeech Computational Paralinguistic Challenge set from the collected audio files using the OpenSMILE algorithm (Eyben et al., 2016; Eyben, Wöllmer, & Schuller, 2010; Schuller et al., 2016). In contrast to the on-device feature-extraction approach in Study 1, those features were extracted from the raw recorded audio files after data collection in Study 2.

We transcribed all raw audio records using the Google Speech-to-text API (version 1). Automatic speech-to-text technology has been shown to be well-suited for transcription tasks in psychological research (Pfeifer, Chilton, Grilli, & Mehl, 2024). The Google API also assigns a sentiment score (*M* = 0.03, *SD* = 0.30) within the interval of [-1; 1] to each speech transcript. Then, we extracted state-of-the-art word embeddings from speech transcripts using the *text* R package (Kjell, Giorgi, & Schwartz, 2021). Word embeddings are vector representations of words in a high-dimensional space, which capture their contextualized meaning and relationships with other words. Specifically, we used the 1,024 dimensions from the second to last layer (layer 23) from the language model "RoBERTa large" as features as recommended in prior work (Liu et al., 2019; Matero, Hung, & Schwartz, 2022).

Machine learning

In both studies, we trained linear Elastic Net regularized regression models (Zou & Hastie, 2005) and non-linear tree-based Random Forest models (Breiman, 2001; Wright & Ziegler, 2017), and a baseline model on the extracted features for the prediction of self-reported affective experience. The baseline model predicted the respective mean values for affective experience of the respective training set for all cases in a test set. We evaluated model performance using a five times repeated ten-fold cross-validation scheme (Bischl, Mersmann, Trautmann, & Weihs, 2012) and used blocking of participants in the resampling procedure to ensure that for each train/test set pair the given participant is either in the training set or in the test set. In the results section, we report on the average (median) model performance across those 50 models. Before predictive modeling, we replaced extreme outliers (mean +/- four times *SD*) with missing values that were imputed as part of the resampling procedure.

In our main analyses, we predicted self-reported affective states (Study 1: arousal and valence; Study 2: arousal, contentment, sadness) from prosodic features (eGeMAPS feature set). For Study 2, we also used semantic content, captured by word embeddings, as predictor. We conducted a number of supplementary analyses that can be found in the repository: As done in prior research (Weidman et al., 2020), we predicted affective states from the extensive 2016 Interspeech Computational Paralinguistic Challenge set (Schuller et al., 2016) to investigate whether a larger prosodic feature set would affect prediction performance. Moreover, in line with prior work (Sun, Schwartz, Son, Kern, & Vazire, 2020; Weidman et al., 2020), we predicted momentary fluctuations in affective experience from each individual's baseline (defined as their median response) to explore if those would be also predictable from speech cues. Specifically for Study 1, we predicted self-reported affective states from prosodic features separately for each sentence condition (positive, neutral, negative) to assess whether the sentiment of the read-out sentences would affect prediction performance. Here, we also conducted F-tests to analyze whether models' prediction errors were significantly different across sentence conditions. For Study 2, we predicted self-reported affective states from a combination of prosodic and semantic features to investigate whether combining those features improved overall performance.

Our prediction models were evaluated based on how accurate new (unseen) samples can be predicted. Throughout this manuscript, we report on the Spearman correlation (ρ) of true and predicted scores and the coefficient of determination (R^2) as measures of model fit. In the repository, we provide the mean absolute error (MAE) and the root mean squared error ($RMSE$) as additional performance measures for prediction models. In our main analyses, we carried out variance-corrected (one-sided) t-tests comparing the R^2 measures of prediction models with those of the baseline models (Nadeau & Bengio, 2003) to determine whether prosodic and semantic features predicted affective experience beyond chance ($alpha$ = 0.05). We adjusted for multiple comparisons (n = 16) via Holm correction.

Our quality checks were data driven, based on voice features (i.e., humans marked as absent when mean voicing score < 0.5, voiced segments per second $= 0$, mean voiced segment length $= 0$). We evaluated the quality of the resulting speech data by training machine learning models on a task they should be able to accomplish if the data are of high quality. Specifically, we trained the models with gender as the dependent variable because past work indicates that speaker gender can be reliably predicted from voice cues (Kwasny & Hemmerling, 2021). Those models predicted speaker gender from prosody with very high accuracy (Study1: Accuracy $_{\text{md}}$ = 91.52%; Study2: Accuracy$_{\text{md}}$ = 95.80%), indicating good data quality.

All data processing and statistical analyses in this work were performed with the statistical software R version 4.0.4 (R Core Team, 2021). For machine learning, we used the *mlr3* framework (Lang et al., 2019). Specifically, we used the *glmnet* (Friedman, Hastie, & Tibshirani, 2010) and *ranger* (Wright & Ziegler, 2017) packages to fit machine learning models. We provide the R code, figures, and results in the project's repository on the Open Science Framework (https://osf.io/a5db8/?view_only=d881fb22ab3340c4b4faa5ca4079db90). Raw data of the voice samples may contain personally identifiable information and, therefore, cannot be shared publicly. We preregistered Study 1 as a transparent account of our work and extended the analytical approach to the data from Study 2.

References

Aishwarya, N., Kaur, K., & Seemakurthy, K. (2024). A computationally efficient speech emotion recognition system employing machine learning classifiers and ensemble learning. *International Journal of Speech Technology*, *27*(1), 239–254. https://doi.org/10.1007/s10772-024-10095-8

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

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). https://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. https://doi.org/10.1037/a0025827

Ben-David, B., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. 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. https://doi.org/10.1044/2015_JSLHR-H-14-0323

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. *Evolutionary Computation*, *20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brooks, J. A., Tzirakis, P., Baird, A., Kim, L., Opara, M., Fang, X., . . . Cowen, A. S. (2023).

Deep learning reveals what vocal bursts express in different cultures. *Nature Human Behaviour*, *7*(2), 240–250. https://doi.org/10.1038/s41562-022-01489-2

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. https://doi.org/10.2196/31724

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382. https://doi.org/10.1038/s41562-019-0533-6

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. *9th International Conference on Speech Prosody 2018*, 714–718. ISCA. https://doi.org/10.21437/SpeechProsody.2018-145

Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, *1*(1), 49–98. https://doi.org/10.1515/semi.1969.1.1.49

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Eyben, F., Scherer, K. R., Schuller, B., 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. https://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. *Proceedings of the International Conference on Multimedia - MM '10*, 1459. Firenze, Italy: ACM Press. https://doi.org/10.1145/1873951.1874246

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

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

große Deters, F., & Schoedel, R. (2024). Keep on scrolling? Using intensive longitudinal smartphone sensing data to assess how everyday smartphone usage behaviors are related to well-being. *Computers in Human Behavior*, *150*, 107977. https://doi.org/10.1016/j.chb.2023.107977

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61-83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

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. https://doi.org/10.1016/j.jbusres.2020.09.020

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

Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *The American Psychologist*, *72*(7), 644–654. https://doi.org/10.1037/amp0000147

Kwasny, D., & Hemmerling, D. (2021). Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks. *Sensors (Basel, Switzerland)*, *21*(14), 4785. https://doi.org/10.3390/s21144785

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. https://doi.org/10.21105/joss.01903

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions. *Perspectives on Psychological Science*, 17456916231217722. https://doi.org/10.1177/17456916231217722

Laukka, P., & Elfenbein, H. A. (2021). Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis. *Emotion Review*, *13*(1), 3–11. https://doi.org/10.1177/1754073919897295

Lim, N. (2016). Cultural differences in emotion: Differences in emotional arousal level between the East and the West. *Integrative Medicine Research*, *5*(2), 105–109. https://doi.org/10.1016/j.imr.2016.03.004

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. https://doi.org/10.1044/2020_JSLHR-19-00258

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

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. https://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. *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 89–94. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wassa-1.9

Matz, S. C., & Netzer, O. (2017). Using Big Data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences*, *18*, 7–12. https://doi.org/10.1016/j.cobeha.2017.05.009

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. https://doi.org/10.1177/0963721416680611

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. https://doi.org/10.3389/fdgth.2022.886615

Nadeau, C., & Bengio, Y. (2003). Inference for the Generalization Error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Pfeifer, V. A., Chilton, T. D., Grilli, M. D., & Mehl, M. R. (2024). How ready is speech-to-text for psychological language research? Evaluating the validity of AI-generated English transcripts for analyzing free-spoken responses in younger and older adults. *Behavior Research Methods*. https://doi.org/10.3758/s13428-024-02440-1

Phan, L. V., Modersitzki, N., Gloystein, K. K., & Müller, S. R. (2023). Mobile Sensing around the Globe: Considerations for Cross-Cultural Research. In L. Tay, S. E. Woo, & T. Behrend (Eds.), *Technology and Measurement around the Globe* (pp. 176–210).

Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009099813.009

Polzehl, T., Schmitt, A., Metze, F., & Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. *Speech Communication*, *53*(9), 1198–1209. https://doi.org/10.1016/j.specom.2011.05.002

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

R Core Team. (2021). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178. https://doi.org/10.1037/h0077714

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

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology.*

Schoedel, R., & Oldemeier, M. (2020). *Basic Protocol: Smartphone Sensing Panel Study.* https://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. https://doi.org/10.1145/3129340

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I–577. https://doi.org/10.1109/ICASSP.2004.1326051

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). https://doi.org/10.21437/Interspeech.2016-129

Seiferth, C., Vogel, L., Aas, B., Brandhorst, I., Carlbring, P., Conzelmann, A., … Löchner, J. (2023). How to e-mental health: A guideline for researchers and practitioners using digital technology in the context of mental health. *Nature Mental Health*, *1*(8), 542–554. https://doi.org/10.1038/s44220-023-00085-1

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. https://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. https://doi.org/10.1037/pspp0000244

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, *7*(3), 386–396. https://doi.org/10.1038/s41562-022-01505-5

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. https://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. https://doi.org/10.1037/emo0000571

Weninger, F., Eyben, F., Schuller, B., 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*. https://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. https://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). https://doi.org/10.1093/gigascience/giab044

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x