scientific reports

Check for updates

OPEN

Multidisciplinary characterization of embarrassment through behavioral and acoustic modeling

Dajana Šipka[1,4]✉, Bogdan Vlasenko[2], Maria Stein[1,3], Thomas Dierks[3], Mathew Magimai-Doss[2] & Yosuke Morishima[3]

Embarrassment is a social emotion that shares many characteristics with social anxiety (SA). Most people experience embarrassment in their daily lives, but it is quite overlooked in research. We characterized embarrassment through an interdisciplinary approach, introducing a behavioral paradigm and applying machine learning approaches, including acoustic analyses. 33 participants wrote about an embarrassing experience and then, without knowing it prior, had to read it out loud to the conductor. Embarrassment was then examined using two different approaches: Firstly, from a subjective view, with self-report measures from the participants. Secondly, from an objective, machine-learning approach, in which trained models tested the robustness of our embarrassment data set (i.e., prediction accuracy), and then described embarrassment in a dimensional (i.e., dimension: valence, arousal, dominance; VAD) and categorical (i.e., comparing embarrassment to other emotional states) way. The subjective rating of embarrassment was increased after participants read their stories out loud, and participants with higher SA scores experienced higher embarrassment than participants with lower SA scores. The state of embarrassment was predicted with 86.4% as the best of the unweighted average recall rates. While the simple VAD dimensional analyses did not differentiate between the state of embarrassment and the references, the complex emotional category analyses characterized embarrassment as closer to boredom, a neutral state, and less of sadness. Combining an effective behavioral paradigm and advanced acoustic modeling, we characterized the emotional state of embarrassment, and the identified characteristics could be used as a biomarker to assess SA.

Embarrassment is a social emotion that most of us experience in daily life, and it occurs when a desired social image of oneself is threatened[1,2]. It can be seen as a self-conscious and, simultaneously, other-conscious emotion: While one is more self-aware in an embarrassing situation, one is also concerned about other people's judgment[2]. Embarrassment is experienced mostly in public with other people. However, it can still occur privately when the audience is imagined and is more likely to occur around strangers than loved ones. Embarrassment is usually accompanied by typical physiological changes, such as blushing[3–5], or changes in the voice and non-verbal behavior, such as avoiding eye contact or lowering one's head. Tangney et al.[2] categorized embarrassment as a "negatively valenced emotion" (pp. 1264), but there are also positive consequences that can arise for the person affected as well as the audience: According to Miller[5], people often react helpfully in embarrassing situations, and showing embarrassment in the form of blushing, for example, elicits a favorable impression of the affected person. Stocks et al.[6] additionally distinguished between personal and empathic embarrassment. Whereas the first is experienced for oneself, the latter is experienced for another person while, for example, observing an embarrassing task. This study, however, focused solely on personal embarrassment.

There are several emotions that are similar to embarrassment, particularly shame, since for a long time, embarrassment was considered to be part of shame. However, embarrassment is viewed as an emotion on its own today and there are distinct differences between two emotions: Shame is related to more moral transgressions, can occur when one is alone, is a more intense emotion and lasts longer than embarrassment. Additionally, embarrassment tends to occur around less familiar people in comparison to shame (see e.g., Refs[1,2,7]). Core aspects of embarrassment, such as the fear of negative evaluation, fear of being rejected by others, and heightened self-consciousness, are also important aspects of social anxiety (SA) as well as social anxiety disorder (SAD). According to Rozen and Aderka[7] embarrassment, SA, and SAD were consistently associated with each other

[1]Department of Clinical Psychology and Psychotherapy, University of Bern, Bern, Switzerland. [2]Idiap Research Institute, Martigny, Switzerland. [3]University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland. [4]Institute for Psychology Clinical Psychology and Psychotherapy Department, University of Bern, Fabrikstrasse 8, Bern 3012, Switzerland. ✉email: dajana.sipka@unibe.ch

in physiological measures, neural activities, and self-reports of emotions for clinical and non-clinical samples. Socially anxious people are generally more easily embarrassed and respond with more intense embarrassment than less socially anxious people[8]. Furthermore, Rozen & Aderka[7] reported in their review that embarrassment has been found to be associated with SA in clinical as well as non-clinical samples over different studies. In both cohorts, people with either higher SA scores or SAD rated social blunders as more embarrassing in comparison to people with lower SA scores respectively without SAD. As with embarrassment, SA occurs mostly in public; if in private, an imagined audience or an imagined reaction from others is necessary. While embarrassment appears sudden, brief, and due to an actual misstep, SA appears gradually and over a longer period of time and can occur without having done anything wrong[9].

Embarrassment is generally neglected in research on basic emotions[10]. Especially when papers report on vocal cues in different emotions, they often examine either the classical basic emotions according to Ekman & Friesen[11] or other emotions than embarrassment (e.g., Refs.[12–15]). Due to the consistent association of embarrassment and SA and the fact that embarrassment is still a neglected emotion, it is important to look more into the emotion itself, its relationship to other emotions, and how it is associated with SA from a basic research and clinical point of view.

Therefore, this paper's main goal was to exploratorily examine embarrassment and capture the emotion from different points of view. On the one hand, embarrassment can be compared categorically to other emotions; it can be described as how it relates to and shares information with them. There are different data sets in different languages, consisting of emotional speech, for which the data have already been labelled and tested accordingly (see e.g.,[16]). On the other hand, embarrassment itself can be described dimensionally in more depth. Grimm et al.[17] proposed a three-dimensional emotion space consisting of the axis valence, arousal, and dominance (VAD), which was also used in this study to describe embarrassment dimensionally.

Previous studies investigating fundamental emotional research used either classical subjective clinical psychology approaches, relying mainly on the participants' self-reports and ratings of a few individual experts, or more objective measures such as neuroimaging methods and psychophysiological measures[18]. If subjective and objective measures are combined in embarrassment studies, they often focus on somatic or neuronal features[3,19] but seldom on voice parameters. The same goes for physiological indicators of SA or embarrassment: Research most often investigated body, hand, and head movements or gaze activity[20]. According to Weeks et al.[21], there are several advantageous characteristics of voice parameters as physiological indicators of, for example, SAD, such as being less biased to subjects' responses and more objective than through self-questionnaires. So, even though participants do not say the same, one can still objectively compare the paralinguistic information of their answers[16].

Human speech can be divided into verbal (linguistic) and non-verbal (paralinguistic) sounds. While it is obvious that verbal sounds play an important role in communication, non-verbal aspects, such as paralinguistics, carry a lot of additional information in conversations, such as the emotional and mental state of the person speaking[22]. Conversely, changes in the voice may indicate changes in a person's mental and emotional state. This study, therefore, combined subjective measures and objective engineering and machine learning approaches to examine embarrassment.

To describe and examine embarrassment from different points of view, this paper had the following four goals: The first goal was to induce embarrassment in participants and test whether the induction was successful and, if so, how embarrassment was related to SA. We hypothesized that embarrassment would indeed be induced, with the participants being significantly more embarrassed during the embarrassment induction task compared to the pre- and post-induction periods. Furthermore, according to previous studies, we also assumed that participants with higher SA scores would get more embarrassed than participants with lower SA scores. Verifying that embarrassment was induced was the only goal with a hypothesis. This was the fundament for the other three following exploratory goals, where we used our acoustic data to gain further insights using automatic speech processing techniques. The second goal was to test how well our trained model could predict our sample data in pre-induction, embarrassment, and post-induction and show the robustness of our embarrassment data. The third goal was to adopt a dimensional approach and map embarrassment onto the VAD dimension. The fourth and last goal was comparing embarrassment to other emotions, thus following a categorical approach. For the third and fourth goals, publicly available emotional speech corpora (i.e., acoustic samples with emotional labels) were used to train our models.

Materials and methods

Participants

Undergraduate psychology students from the University of Berne, Switzerland, were recruited for this study. The average age of the sample ($N = 33$) was 23.73 years ($SD = 4.68$), with 78.8% female participants ($n = 26$). The exclusion criteria were (a) any current or past neurological or psychological disorders, (b) regular medication intake, and (c) regular substance abuse. As the intervention was unexpected for the participants and might have been too intense for people with SAD, we excluded SAD as well as any other possible comorbid disorders. For their participation, they received 1.5 out of 12 mandatory experimental credits. All participants gave written informed consent to participate in the study. The ethics committee of the faculty of human sciences at the University of Bern approved the study protocol (2020-06-00004). The experiment was performed in accordance with the relevant guidelines and regulations according to the Declaration of Helsinki.

Procedure

See Fig. 1 for an overview of the study procedure. If the participants did not fulfil any exclusion criteria, they advanced to fill out the questionnaires Short Form Social Phobia Scale (SPS-6), Short Form Social Interaction

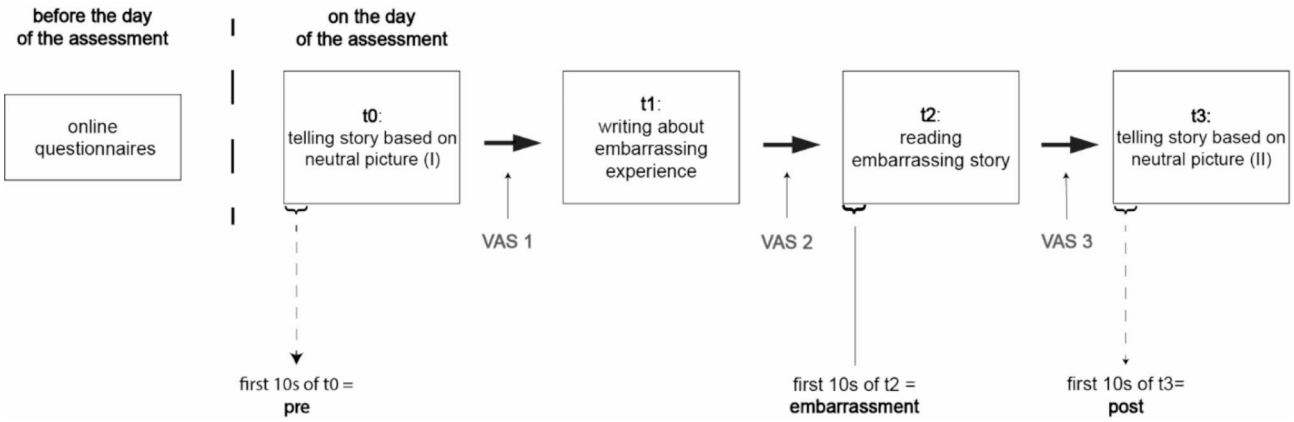

Fig. 1. Overview study procedure. Before starting the main assessment, participants answered online questionnaires. On the day of the assessment, participants were told to tell a story based on a neutral picture before (t0) and after (t3) embarrassment induction (t2). The first ten seconds of speech in t0, t2, t3 were used as pre-embarrassment induction (pre), embarrassment induction, (emb), and post-embarrassment induction (post) for the acoustic analysis. VAS = visual analog scale.

Anxiety Scale (SIAS-6), Community Assessment of Psychic Experiences (CAPE), and demographic data online via Qualtrics not later than 12 h before the assessment appointment.

The whole assessment (t0 – t3) was audio recorded, and the participants were informed prior to the start of the assessment. The experimenter conductor was instructed to respond neutrally respectively to not give either a positive reaction (e.g., laugh along with the participant) or a negative reaction (e.g., look dismissive) during the procedure and had all instructions written on a script to maximize standardization. This study was embedded in a larger study based on Mota et al.[23] and represents part 2 of the procedure. Since part 1 is irrelevant to this study, it was not further explained here (see[23], for the study procedure of part 1). At t0, participants were shown a neutral picture from the picture set of Mota et al.[23] and were instructed to tell a story about the picture for at least 30 s. If they did not meet the 30-second mark, they were asked to talk more about the picture. At t1, they received a short explanation of the characteristics of embarrassment and how it differed from shame (see introduction for differences). The assignment then was to write about a situation where they felt embarrassed, which happened no longer than one year ago. The participants were instructed to write only a few sentences about the setting of the embarrassing situation and then to mainly focus on the following aspects of their experience in the situation: emotions (i.e., "What did you feel in this situation?"), cognitions, physiological reactions, their own behavior, and the behavior of the bystanders. Therewith, participants were more likely to be immersed in the whole experience and not only recollect mere cold facts of the situation[2]. After being informed orally, they also received all instructions in written form. Participants had a maximum of 4 min to come up with a situation and then a maximum of 14 min to write about the situation. At t2, they were asked to read the story aloud to the instructor and were instructed to talk for at least 30 s. If they did not meet the 30-second mark, they were asked to talk more about the situation. At t3, they had the same task as at t0 with the same instructions. Between each time point (t0 – t3), they had to indicate their level of embarrassment on a visual analog scale (VAS).

At the end of the study (after t3), participants were asked to guess the study's purpose. They were then informed about the actual purpose of the study. For ethical reasons, they were asked to state their current well-being, and the instructor intervened if a participant stated particularly low well-being. Additionally, participants were provided with a list of addresses and telephone numbers for psychotherapeutic support if needed later.

Measures

Short form social phobia scale (SPS-6) and short form social interaction anxiety scale (SIAS-6)

The Social Phobia Scale (SPS[24]) and the Social Interaction Anxiety Scale (SIAS[24]) are two self-report questionnaires to measure different aspects of SA. The SPS measures SA in performance-related situations (e.g., fear of attracting attention while queueing), whereas the SIAS measures SA in interactional situations (e.g., difficulty in talking to other people). For this study, the short form of the SPS[25] and the SIAS[25] were used, which were directly translated into German for this study. The two questionnaires are presented together with 6 SPS-6 items and 6 SIAS-6 items rated on a Likert-Scale from 0, "not at all", to 4, "extremely", with a total sum score ranging from 0 to 48. Peters et al.[25] showed that the SPS-6 and SIAS-6 have satisfactory convergent, construct, and criterion validity. Concerning the reliability, the paper from Ouyang et al.[26] showed a satisfactory Cronbach's alpha. According to Peters et al.[25] the cut-off scores for clinically relevant SA symptoms are SPS-6 ≥ 2 and SIAS-6 ≥ 7.

Visual analog scale (VAS)

For the measurement of the strength of embarrassment participants felt before, during, and after the intervention, a paper-pencil visual analog scale (VAS) was used at each time point (i.e., after t0, t1, and t2). Participants marked on a 10 cm long line how embarrassed they felt at the moment from 0, "not embarrassed at all", on the left end to

10, "extremely embarrassed", on the right end. The measured length from the zero point to the participants' mark on each line indicated the strength of embarrassment for each time point (cf. Delgado, 2018[27]). The interrater variability of two independent raters who measured the distance between all lines was $r(97) = 0.998$.

Recording procedure and device

Each participant's assessment was audio-recorded from t0 to t3. The recording device was always placed on the table in front of the participant with approximately 1 m from the edge of the table opposite the participant. The recording device SONY ICD-UX570 was used. The sample rate was at 44.1 kHz with a bit depth of 16-bit. The audio data were stored in "Waveform Audio File Format" (.wav).

Statistical analysis – behavioral data

The statistical analysis of the behavioral data was performed with R Studio (Version: 2024.04.2 + 764;[28]). Descriptive statistics were reported, and to examine differences and relationships, non-parametric tests were used due to a lack of normal distribution in the behavioral data. Linear mixed models (LMMs) were conducted to test the influence of gender.

Acoustic data analysis

The acoustic data from each participant's first 10 s at the beginning of speech in t0, t2, and t3 phases were used for pre-embarrassment induction (pre), embarrassment induction (emb), and post-embarrassment induction (post), respectively. A fixed length of initial 10 s was chosen for the following two reasons: Firstly, from a behavioral perspective, participants were likely to experience the strongest embarrassed at the beginning of their speech. Over time, embarrassment may decrease due to habituation, resulting to greater variability across participants. By using a fixed length, the influence of habituation could be minimized. Secondly, from a machine learning perspective, fixed-length acoustic segments are more suitable for training predictive models in paralinguistic classification tasks. The pipelines for acoustic analysis used in this study are publicly available in the following GitHub repository (https://github.com/idiap/embarrassment_acoustic/).

Auxiliary emotional corpora

To train regression and classification models for the conceptual and dimensional description of embarrassment, two well-known emotional datasets in German were selected: the Berlin Emotional Speech Database (EMO-DB)[16] and the "Vera am Mittag" (VAM) dataset[29].

The VAM corpora consist of 947 emotional German speech samples collected from 47 speakers (36 female). Speech segments were selected from 12 broadcasts of the TV talk show "Vera am Mittag" (in Engl.: "Vera at noon"). The weighted average values with evaluator weighted estimator (EWE) techniques of VAD emotional dimensions were used to train our regression models for the dimensional VAD representation of embarrassment[17]. Each speech sample in the database has EWE smoothed VAD dimensional labels in the range of [-1,+1]: valence (negative: -1 & positive: 1), arousal (calm: -1 & excited: 1), and dominance (weak: -1 & strong: 1)[30].

EMO-DB covers 7 emotions (i.e., anger, joy, neutral, sadness, disgust, fear, and boredom). The corpus consists of 10 professional actors (5 female) speaking out 10 predefined phonetically balanced and emotionally neutral sentences. We utilized a subset of 493 utterances, which achieved naturalness and recognizability rates of 60% and 80%, respectively, as obtained during a perception test involving 20 subjects. This subset was used to train our emotion categories-based classification models.

Acoustic feature representations

Considering a comparably sparse amount of training samples for acoustic modeling, we decided to use knowledge-based handcrafted features and pre-trained data-driven feature representations. In the employed acoustic feature representations, we do not use phonetic-level information. Handcrafted features are also interpretable. Thus, it can help in gaining insight into the acoustic variations related to speech production in the present study's context.

For the knowledge-based *handcrafted* feature representations (FRs), we used the (a) ComPaRE 2016 feature set provided by the openSMILE extraction tool (see[31] for more information), which has been studied for several paralinguistic speech processing tasks[32]. The ComPaRE set contains 6373 static turn-level features resulting from the computation of functionals (statistics) over low-level descriptor contours.

Considering top performance positions on challenge leaderboards and state-of-the-art neural embeddings for the Speech Emotion Recognition (SER) task from the Speech processing Universal PERformance Benchmark (SUPERB) challenge[33], we employed (b) Wav2Vec2 (WV2 EM)[34], fine-tuned for dimensional SER on the MSP-Podcast[35] as emotion data-driven feature representation, and (c) WAVLM (large)[36] as general data-driven feature representation.

Machine learning experimental setup

To address our second goal (i.e., test how well our classification model could predict speaker state and distinguish between non-embarrassment and endorsement data samples), we used a speaker-independent experimental protocol, more precisely leave-one-speaker-out (LOSO) protocol. This protocol simulates realistic conditions and maintains a good balance between training and testing subsets. In the LOSO protocol, we used speech samples from 32 speakers to train predictive models, and speech samples from the one remaining speaker were used for testing. This procedure was repeated 33 times, and predictions were aggregated into one file to estimate recognition rates. To measure recognition performance, we used unweighted average recall (UAR). During the experimental study, we used both three-class and two-class experimental settings. The three-class

setting included pre-embarrassment, embarrassment, and post-embarrassment labels. In the two-class setting, we provided a selection of different combinations or grouped pre- and post-embarrassment classes into one class (i.e., non-embarrassment). Random Forest (RF) and Support Vector Machines (SVM) classifiers were used to train predictive models that take ComPaRE FR as input. The results obtained from the speaker-independent evaluation study for the second goal is presented.

For the third goal (i.e., adopting a dimensional approach and mapping embarrassment onto the VAD dimension), we employed a cross-corpora approach, where first three independent Random Forest-based regressors that take as input ComPaRE FR and predict valence, arousal, and dominance values, respectively, were trained on the emotional speech samples of the VAM dataset. As mentioned earlier in Sect. *Auxiliary emotional corpora*, the prediction values for each of the VAD emotional dimensions are lie in the range [-1,1]. After training the regressors on the VAM database, we used the models to predict VAD labels and estimate a possible location of the embarrassment state in the circumplex model of affect introduced by Russell (1980). For this purpose, we used pre- and post-embarrassment samples as a reference and modeled possible shifts of numerical VAD labels in the context of these reference samples. The results of this experimental study are presented.

For the fourth goal (i.e., comparing embarrassment to other emotions, following a categorical approach), as there are no categorical emotion labels associated with the data collected in our study, we again employed a cross-corpora approach. In this approach, a categorical emotion classifier was first built on EMO-DB corpus (see Sect. *Auxiliary emotional corpora*) and then the classifier outputs for our speech data were analyzed. Considering the different recording conditions and the phonetic differences between High German (in EMO-DB corpus) and Swiss German (in our data), for robust analysis, we took a multiple classifier/expert approach. As illustrated in Fig. 2, we trained three different categorical emotion classifier systems based on (a) handcrafted FR, (b) data-driven feature representation tuned for emotional analysis task (emotion data-driven FR), and (c) general data-driven FR on EMO-DB corpus. Each of the emotion classification systems were trained to predict class conditional probabilities of seven emotion classes (including neutral class). For our analysis, we passed each

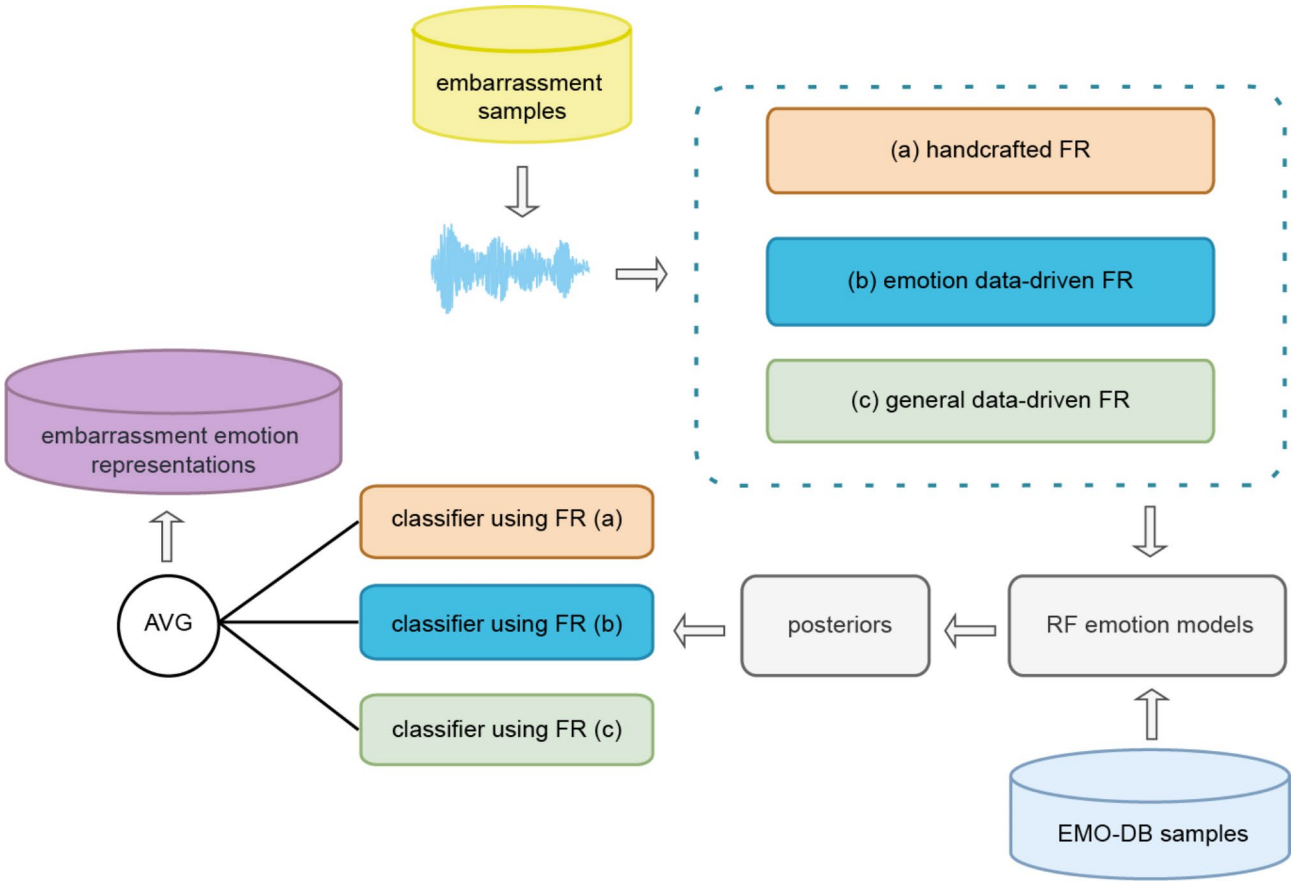

Fig. 2. Flow chart of the emotional categorical analysis with EMO-DB samples. The presented pipeline has been used to represent pre-embarrassment, post-embarrassment, and embarrassment samples with aggregated class-conditional probabilities (referred to as posterior-based emotional representations) from categorical emotional models trained on EMO-DB. The proposed multiple classifier/expert approach is based on three expert systems trained with handcrafted FR, emotional data-driven FR, and general data-driven FR. FR = Feature Representation. RF = Random Forest. EMO-DB = a database (see[16] for more information). AVG = average. For (**a**–**c**), see Sect. *Acoustic feature representations*.

of our speech data samples through those three trained classifiers and took the average of the emotion class conditional probabilities estimated by the three classifiers. The resulting average emotion class conditional probabilities were then analyzed. As in the previous cases, we used pre- and post-embarrassment samples as reference samples and treated them as separate classes. The results of this experimental study are presented in Sect. *Categorical model of embarrassment using the 7emotions from EMO-DB*.

Rate of speech analysis

To compute rate of speech, we adopted graphemes-per-second as a metric, because the grapheme-to-phoneme relation in German is shallower than the relation in English. In other words, the rate of speech in German is well represented by graphemes. We used OpenAI's Whisper model (https://huggingface.co/openai/whisper-large-v3) available from Huggingface hub to transcribe the speech data. Then, we removed all punctuation marks and blank spaces in the transcription and counted the number of graphemes/characters. The grapheme count was finally divided that by 10 s to obtain the grapheme-per-second measure.

Results

SPS-6 & SIAS-6 results and embarrassment induction

Table 1 shows the results from the questionnaires SPS-6 and SIAS-6. 16 Participants were above the SIAS cut-off (i.e., sum SIAS ≥ 7), 17 participants were above the SPS cut-off (i.e., sum SPS ≥ 2), and 9 participants were above both cut-offs (i.e., sum SIAS ≥ 7 and sum SPS ≥ 2). Therefore, according to the cut-offs, 27.3% (9 / 33) showed clinically relevant SA symptoms in both questionnaires. This rate is consistent with a previous epidemiological and cross-cultural study on screening of SA (pp. 10: "23–58% across the different countries")[37].

The first goal of the study was to verify the embarrassment induction. The Shapiro-tests for the differences of the 3 VAS scores (VAS 1 – VAS 2, VAS 1 – VAS 3, VAS 2 – VAS 3) (see Fig. 1) between t0 and t3 showed a significant result for the difference VAS 2 – VAS 3 ($p = .023$). Therefore, all further calculations were made with non-parametric tests. Multiple pairwise Wilcoxon tests showed a significant difference between the VAS 2 ($M = 2.52$, $SD = 2.09$) and VAS 3 ($M = 4.40$, $SD = 2.38$) score ($V = 551$, $p < .001$). This means that the participants got significantly more embarrassed after reading the story aloud to the conductor than after only writing the story. This can also be observed visually in Fig. 3 for 31 out of 33 participants. The other differences were not significant (VAS1 – VAS2, $p = .13$, VAS 1 – VAS 3, $p = .20$). To account for the gender disbalance ($n = 26$ females) in the significant differences between VAS 2 and VAS 3, we used LMMs. First, the time only model was calculated with time (VAS 2, VAS 3) as a predictor and ID as a random effect (i.e., accounting for the fact that all participants have different baseline values). As expected, the time only model was significant ($b = 1.88$, $SE = 0.26$, $t(32) = 7.22$, $p < .001$). Then the covariate gender was added to the model (time*gender). The interaction between time and gender was not significant ($b = 0.57$, $SE = 0.64$, $t(31) = 7.22$, $p = .89$), implying that the effect of time on the VAS values did not differ by gender.

Spearman correlations were calculated between the SIAS-6 and SPS-6 total scores, the total sum of both questionnaires and the VAS 2 – VAS 3 difference score. There was a significant negative correlation between the difference score and the total sum score of the SPS-6 and SIAS-6 ($r_s = -0.36$, $p = .037$). This means that the higher SA is (here defined by the overall sum of both questionnaires), the more embarrassed a person felt after reading the story out loud.

Prediction of embarrassment based on acoustic features

The second goal of the study was to test the possible prediction performance of machine learning models applied to predict embarrassment. During our preliminary analysis of embarrassment prediction, we used three-class and two-class configurations for classification experimental setups. In the two-class settings, in addition to the non-embarrassment class concept (see Sect. *Machine learning experimental setup*), we used a pre- vs. post-embarrassment configuration. Additionally, considering the scarcity of collected embarrassment data and the need for better interpretability, we used only knowledge-based acoustic features, namely the ComPaRE (see Sect. *Acoustic feature representations*) feature set.

Table 2 shows the prediction performance (i.e., the probability of predicting the correct set) of RF and SVM. The best performance was found for predicting the pre-set vs. the embarrassment-set (pre vs. emb), where the right prediction for the respective set was made in 84.8% of the cases for RF and 86.4% for the SVM (since it was a two-class configuration, a performance by chance would be 0.5. The results here indicated a very good prediction performance). The second- and third-best performance was found for the prediction of the post vs. embarrassment (post vs. emb) and for the pre- and post-set combined vs. the embarrassment-set ((pre + post) vs. emb). The pre- vs. post-comparison was slightly above chance and could not be discriminated satisfactorily. In the case of the three-class configuration (pre vs. emb. vs. post), the main misclassification confusion was observed between the pre- and post-embarrassment classes: The pre- and post-state showed a tendency to have

Questionnaires	*N*	*M*	*SD*	*Mdn*
SPS-6	33	2.52	3.01	2.00
SIAS-6	33	7.00	4.02	6.00
S (SPS–6 & SIAS–6)	33	4.76	4.18	4.00

Table 1. SPS-6 & SIAS-6 descriptive statistics. SPS-6: Short form Social Phobia Scale; SIAS: Short form Social Interaction Anxiety Scale.

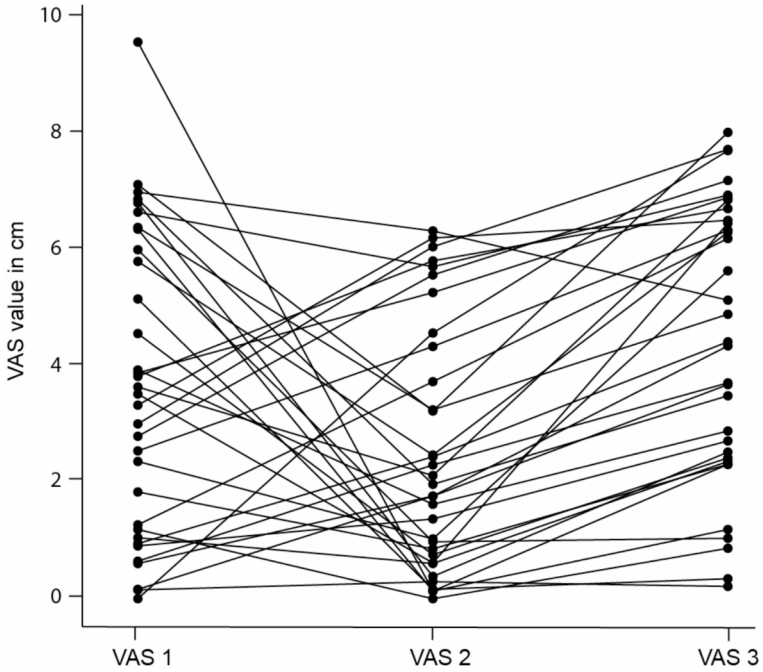

Fig. 3. VAS values connected over all time-points for each participant. VAS-1 measured embarrassment between telling a story based on a neutral picture and writing about an embarrassing experience. VAS-2 measured embarrassment between writing about an embarrassing experience and reading the embarrassing story. VAS-3 measured embarrassment between reading the embarrassing story and telling a story based on another neutral picture. There was a significant difference between VAS-2 and VAS-3 ($V = 551$, $p < .001$), where 31 out of 33 got significantly more embarrassed after reading the embarrassing story.

	Pre vs. emb	Post vs.emb	Pre vs. post	(Pre + post) vs. emb	Pre vs. emb vs. post
RF	**0.848**	**0.833**	0.621	**0.803**	0.636
SVM	**0.864**	**0.818**	0.591	**0.818**	0.596

Table 2. Prediction performance UAR of RF and SVM for various combinations of pre-, embarrassment- and post-sets. UAR = Unweighted Average Recall. RF = Random Forest. SVM = Support Vector Machine. emb = embarrassment. The bold print indicates trends for a higher prediction accuracy for RF and SVM.

similar characteristics since their predictive performance was slightly above chance (see Sect. *Categorical model of embarrassment using the 7 emotions from EMO-DB* for quantitative proof of similarity).

Supplementary Table 2 shows the top-ranked features for each combination of prediction from Table 2. Feature ranking was conducted based on RF feature importance rates. The glossary in Supplementary Table 1 can be consulted for more information on the features. A more detailed description of acoustic features and their corresponding mathematical models can also be found in the paper of Eyben[38]. Since the best prediction performances were found for pre vs. embarrassment and (pre + post) vs. embarrassment, only their top-ranking features were examined in more detail. For selected set configurations (pre vs. embarrassment and (pre + post) vs. embarrassment), mainly vocal tract-related dynamic features were presented in the top-ranking list (i.e., the most discriminative features). Most of the top-ranked acoustic features for the (pre + post) vs. embarrassment task represent temporal dynamics of envelope of short-term spectrum which relates to modulation frequency information, indicating changes in energy in frequency bands across time due to speech articulation (movement of jaw, tongue and lips)[39].

One way to ascertain whether modulation frequency information is indeed playing a central role here is to compare speaking rate, as changes in rate of speech leads to changes in modulation spectrum[40]. More precisely, peak modulation frequency tends to correlate with speech rate. We used the speech recognition system to transcribe the pre, post, and embarrassment speech and calculated speaking rate in terms of graphemes-per-second for each condition, respectively, each time point. As mentioned earlier, graphemes-per-second is a good indicator of speaking rate in our case, as in German the grapheme-to-phoneme relation is shallow when compared to English. Figure 4 shows the distribution of grapheme-per-second for 33 speech samples (33 speakers) for each condition. We found that the speaking rate was higher for embarrassment (mean: 11.92) than for pre-embarrassment (mean: 7.67) and post-embarrassment (mean: 8.16), suggesting that modulation frequency information indeed helps to distinguish state of embarrassment.

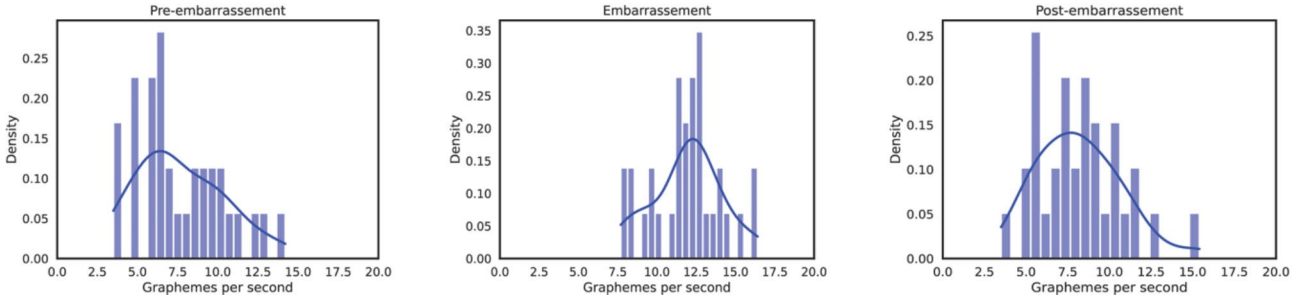

Fig. 4. Histogram and Gaussian kernel density estimation for grapheme-per-second rates for pre-embarrassment (left), embarrassment (middle), and post-embarrassment (right).

Dimensional modeling of embarrassment using the dimensions Valence, arousal, and dominance (VAD) from VAM

In order to find the possible location of the embarrassment state in the circumplex model of affect introduced by Russell[41], we decided to process the embarrassment data using pre-trained VAD regressors. Figure 5a and c shows a density map for the predicted valence, arousal, and dominance levels for pre vs. embarrassment and post vs. embarrassment states for all 33 participants (represented as dots). Considering the sparsity of the collected data, we used knowledge-based feature representations (i.e., ComParE) for acoustic modeling in our regression models.

Predictive models were trained on the VAM database speech samples. The regressors trained on VAD emotional dimensionalities were used to predict VAD levels for pre-, post-, and embarrassment samples. To evaluate changes in VAD levels for embarrassment phenomena, we mapped the obtained predictions in 2D plots presented in Fig. 5a and c.

Arousal (Fig. 5b) & dominance (Fig. 5c) seem to indicate embarrassment phenomena (more points are on the embarrassment side from the diagonal), while valence (Fig. 5a) does not show any predictive value (dots are quite evenly distributed around the diagonal) and should not be used without further linguistic post-hoc tests.

Nevertheless, none of the VAD dimensions could differentiate enough between pre- vs. embarrassment and between post- vs. embarrassment states. The obtained results showed that embarrassment phenomena could be characterized by comparatively complex changes in VAD levels compared to pre vs. embarrassment and post vs. embarrassment states.

Supplementary Table 3, the mutual information (MI) between the regression task for the VAD emotional dimensions on VAM samples and the embarrassment detection task can be found. Selected acoustic features have high MI rates for both tasks (embarrassment detection and VAM-based emotion modeling). As in our previous feature ranking analysis based on RF feature importance, vocal tract-related modulation features are influential in the selected features with high shared MI. Additionally, Supplementary Table 4 shows MI between EMO-DB and embarrassment.

Categorical modeling of embarrassment using the 7 emotions from EMO-DB

For the classification task, the database EMO-DB was used to train the classifiers for acoustic emotion categories. The database contains emotional speech data from 5 female and 5 male actors speaking in German. They portrayed the following 7 emotions: happiness, sadness, disgust, fear, boredom, anger, and neutral state, to which our embarrassment data set was compared.

Figure 6a shows the probability density functions (PDF) for posterior probability of all EMO-BD emotions over the 3 audio time points (pre, embarrassment, post). In other words, it shows the probability of how much each emotion was represented in our embarrassment sample for each time point. There is a skewed distribution for each time point and a clear distribution change in posterior probabilities of the emotions boredom, sadness, and neutral state over the time points. The most indicative changes in these 3 emotions are plotted in Fig. 6b. The histogram plots show average posteriors for sadness, boredom, and neutral state.

For quantitative proof of the significance of emotional posterior modeling, we used a t-test to evaluate the average posteriors of each emotional state. The above-mentioned indicative changes were also observed in Table 3 with the mean posterior values of each emotion: The obtained p-values for sadness, boredom, and neutral state posteriors are highly significant for both pre vs. embarrassment comparisons and the post vs. embarrassment comparisons. At the same time, the p-values for pre vs. post comparisons for all emotional states are not significant. With that, it could be shown that pre- and post-states show indeed similar characteristics, which was already assumed in Table 2.

In conclusion, being embarrassed seems to shift the voice toward less sadness but more boredom and neutral state-related characteristics. But this shift is only temporary and disappears after a while (at post).

Supplementary Tables 3 and 4 shows the MI for categorical emotion classification on EMO-DB and the dimensional modeling. We used the complete emotion set presented in the EMO-DB database and binary classification mapping configurations: high/low and arousal/valence. Results in Supplementary Tables 3 and 4 show that the embarrassment phenomenon could cause changes in both emotional dimensions, such as arousal and valence. Hence, the highest shared MI in the emotion classification task with the 7 emotional classes experimental setup can be observed.

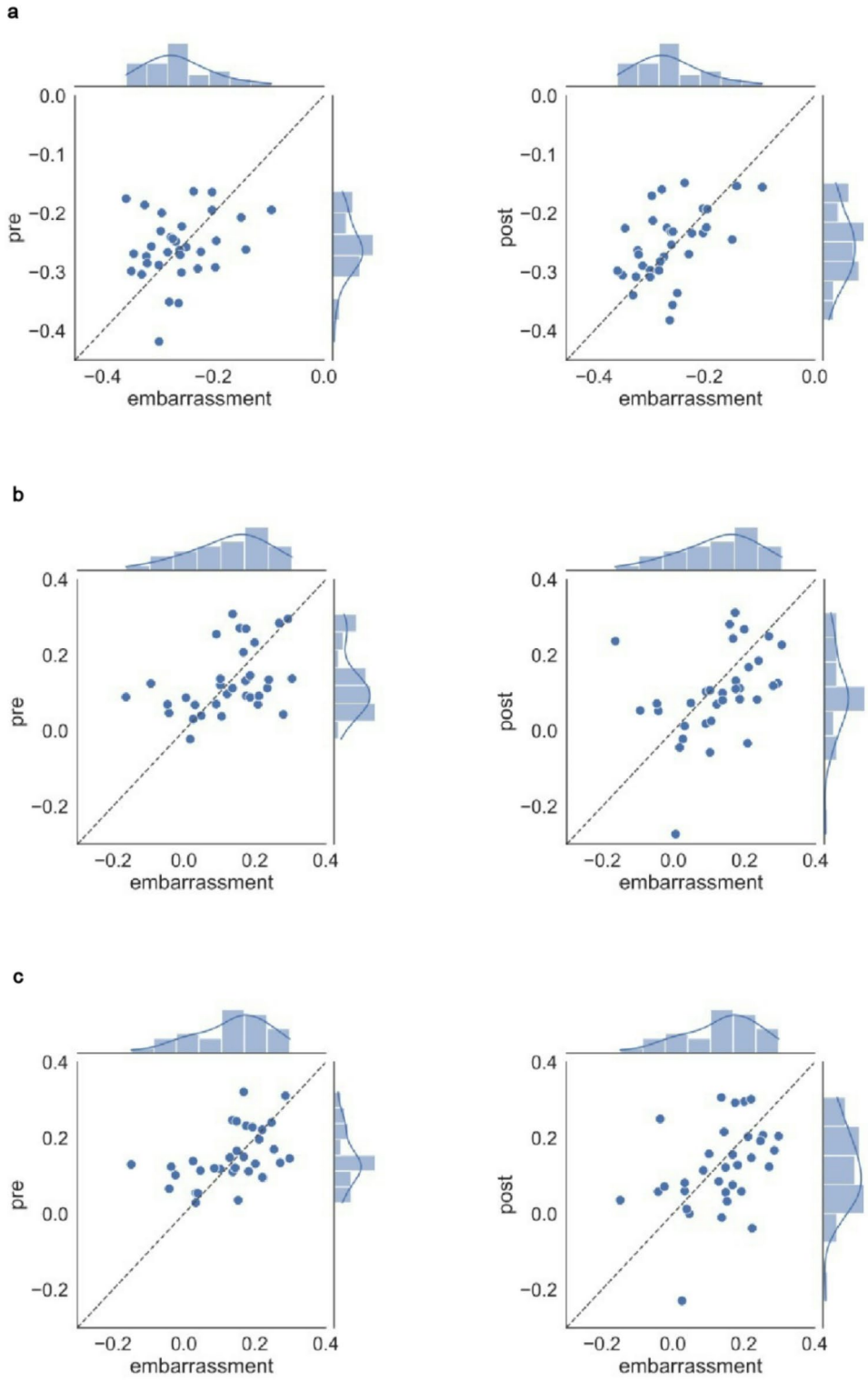

Fig. 5. Predicted levels of (**a**) valence, (**b**) arousal, and (**c**) dominance. Scatter and density plot for pre vs. embarrassment and post vs. embarrassment states. The 0-point indicates a neutral state.

Discussion

This study aimed to systematically examine embarrassment categorically and dimensionally with an interdisciplinary self-report and machine-learning approach.

In the first step, the induction was tested with subjective self-reported measures (i.e., SPS-6, SIAS-6, VAS), and it was shown that participants became significantly more embarrassed after reading the story aloud than

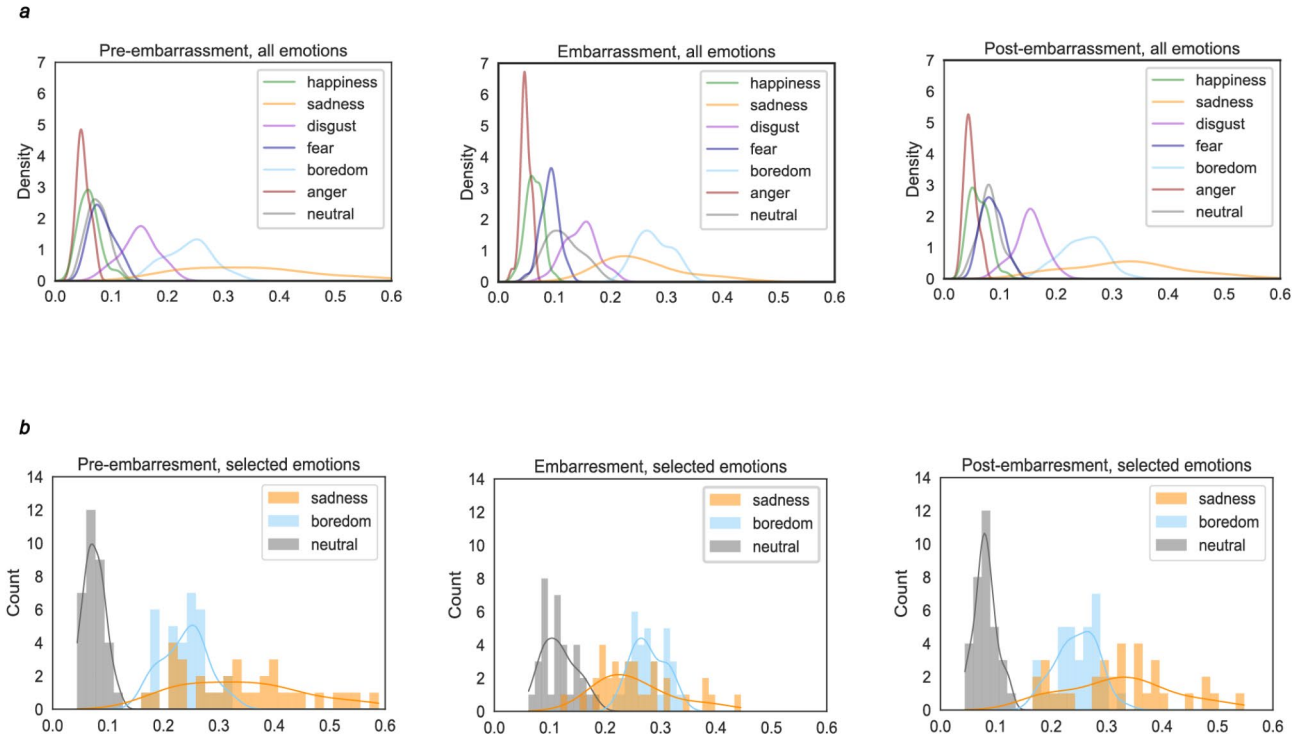

Fig. 6. The density x-axis shows the average probability for the observed emotional class over all participants (sum of all probabilities is 1). Count = number of participants. (**a**) The line plots show the distribution approximations of posterior-based seven emotional states for pre-embarrassment (left), embarrassment samples (middle), and post-embarrassment (right). (**b**) Histogram and line plots of selected three emotional states, showing significant changes between pre-embarrassment and embarrassment, for pre-embarrassment (left), embarrassment samples (middle), and post-embarrassment (right).

Emotion	Mean			*p*-value		
	Pre	Emb.	Post	Pre vs. emb.	post vs. emb.	Pre vs. post
Happiness	0.06	0.06	0.06	0.503	0.652	0.818
Sadness	0.34	0.25	0.33	< 0.001	0.001	0.506
Disgust	0.15	0.15	0.15	0.699	0.397	0.697
Fear	0.08	0.09	0.09	0.079	0.146	0.685
Boredom	0.24	0.28	0.25	< 0.001	< 0.001	0.402
Anger	0.05	0.05	0.05	0.574	0.582	0.317
Neutral	0.08	0.12	0.08	< 0.001	< 0.001	0.200

Table 3. EMODB's emotional representations in mean posterior values over all subjects ($N = 33$) and the p-value for the t-test for different pairs of mean. *p*-values were retrieved from *t*-tests.

only writing it and that the embarrassment induction was successful. This was particularly the case for people who reported higher SA symptoms before the experiment.

In the second step, the robustness of the embarrassment data was tested based on acoustic features. The robustness of the data set was shown through high prediction performance trends for the comparisons pre vs. embarrassment and pre- and post-set combined vs. the embarrassment-set, with SVM showing slightly better predictions than RF. The pre- vs. post-prediction performance was slightly above chance; one hypothesis is that people return to the initial state so that pre and post cannot be distinguished from each other.

In the third step, embarrassment was examined dimensionally in the VAD dimensions. It was shown that there was a shift in the dimensions of arousal and dominance, while there was none for valence. In order to obtain an applicable prediction with valence, one would have to adapt linguistic-based techniques since embarrassment is a complex emotion, and valence is dependent on different factors (as is natural language processing in general domain-dependent). A participant could, for example, try to mask their embarrassment by altering the voice so that valence would be perceived as positive, even though the content of the audio data is negative. These results suggest either that embarrassment might not be described with simply VAD dimensions since there were no

differences detected (i.e., distribution graphs were comparable and comparable number of points at each side of the diagonal), or that the model itself is too simple to capture a complex emotion like embarrassment.

In the fourth step, embarrassment was examined categorically by comparing it to seven other emotional states. It was shown that the state of embarrassment had characteristics of less sadness and more boredom and a neutral state in comparison to the pre- and post-state. On the other hand, the pre- and post-states did not show any significant differences, implying that the participants returned, after feeling embarrassed, to a pre-similar state. It additionally shows that our chosen method of embarrassment induction only leads to a temporary heightened embarrassment state and might, therefore, be safe to use with subclinical or clinical samples. This generally suggests that embarrassment is indeed a complex emotion since we observed a shift in multiple emotions (see Fig. 6a). There might also be cultural differences that might explain why there is a higher posterior probability of sadness at pre (one would rather expect a neutral state) or due to the emotions being acted and therefore over-exaggerated.

The behavioral data demonstrated that reading the story out loud induced embarrassment (in comparison to writing the embarrassing story down). An advantage of this individualized embarrassment induction compared to other induction methods, like singing or holding a presentation, is that writing a personal story and reading it out loud makes sure that it is embarrassing for each person individually, while singing or holding a presentation might even be pleasant for certain people. Indeed, 31 out of 33 participants reported increased embarrassment after speaking about their embarrassing experience. In addition, the machine learning models found shifts for embarrassment in acoustic features, arousal and dominance on the VAD dimensionality, and emotional characteristics of sadness, boredom and neutral state. One explanation is that the articulation changes when people get embarrassed. They might try to become more in control of their voice and to articulate more precisely, hence the change in dominance and the heightened boredom and neutral state (which could also explain the decrease in sadness). Another explanation could be that participants emotionally distanced themselves from the situation due to embarrassment and, therefore, the characteristics of the voice changes. It would have been interesting to have an additional VAS at the end of t3 to test whether the subjective feeling of embarrassment would coincide with the return of the voice to the pre-similar state.

In general, embarrassment can be predicted with high accuracy from the voice only, even though we had a small sample and fixed durations of voice samples (i.e., 10 s, respectively, the same amount of information for each participant), which shows that our fundamental study set-up is valid. It also shows the overall good robustness of the data set, that the classification models trained on EMO-DB can be used to understand the embarrassment phenomenon, and that the t-tests provide an additional source of information for the classification models.

The top-ranking FRs in acoustic analysis has identified two aspects in the LLDs that stand out. First, the FRs are either based on auditory spectrum or based on MFCCs. Both these FRs tend to parameterize the spectral envelop of the short-term spectrum, which carries information more related to the acoustic changes due to the change in the shape of the vocal tract system when producing speech[42–44]. Second, in all the cases the LLDs were capturing the temporal dynamics of the envelope of short-term spectrum using either delta feature computation (denoted as "de" in the FR definition) or RASTA processing (denoted as "Rasta" in FR definition). Taken together, these two aspects indicate that the classification systems are focusing on modulation frequency information[39]. Consistent with the changes of modulation frequency information, we further found that rate of speech was increased during embarrassment. Another advantage of our setting is the high degree of naturalness, respectively, the ecological validity since emotion-related studies often work with acted respectively simulated emotions[45,46].

Our study shows the following limitations: Firstly, no control or comparable group existed. Due to the within-study design, every participant is a control for themselves for intra-individual differences, but there is no comparison for inter-individual differences for time-effects. One could, for example, add another group that writes about a neutral, instead of an embarrassing, story. Secondly, even though we were able to find robust effects, the number of participants ($N = 33$) in the study was rather small. Thirdly, the majority of the participants were female ($n = 26$). Therefore, we were not able to control for gender differences. However, the LMMs implied that there was still a significant difference for the behavioral data when controlling for gender. Fourthly, we did not control for a linguistic bias. This was already acknowledged in the results part, where valence was not further considered. Fifthly, as no data set exists which could be used to train our models in Swiss German, they were trained in High German. Sixthly, carry-over effects due to part 1 of the study cannot be ruled out. It could be, for example, that participants were more sad or bored due to part 1.Lastly, the participants knew that they were recorded, which could have influenced the expression of embarrassment. However, it remains an open question, if there was any influence, whether it led to a more controlled or exaggerated expression of embarrassment.

By examining paralinguistic voice parameters associated with embarrassment, this paper contributes to multiple research gaps at once. Due to the especially big gap in voice analyses for embarrassment, this paper focused mostly on the fundamental description and classification of embarrassment with voice parameters rather than the association between SA and embarrassment. There was a study conducted by Simon-Thomas et al.[10] which examined vocal bursts, however, to our knowledge, this is the first study using acoustic analyses to study embarrassment in a multidisciplinary approach.

A future paper will address this association between a healthy sample (control group) and a SAD sample. The SAD sample would be particularly interesting to investigate since we found a correlation between SA and the height of embarrassment, assuming at least more robust and larger effects. From a clinical point of view, it would additionally be interesting to ask the participants with SAD in anxiety inducing social situations whether they notice changes in their voice and if they think that others might notice it, since one of the key processes in people with SAD in anxiety-inducing social situations is the heightened inward attention on themselves and their interoceptive information and with that the fear that others might notice their anxiety[47]. Additionally, due to the potential negative consequences of experiencing repeated embarrassment (cf[48]). as well as embarrassment

being part of psychiatric diagnoses (e.g., in SAD, Taijin kyofusho, agoraphobia) or consequences of psychiatric disorder (e.g., in Trichotillomania, excoriation disorder) (Diagnostic and Statistical Manual of Mental Disorders 5th edition (text rev.);[49]). Future intervention studies could focus on how to cope with embarrassment. For example, by providing psychoeducation on embarrassment (cf[50]). or practicing cognitive restructuring (both typical components of a cognitive behavioral therapy) for embarrassing and SA inducing situations. Furthermore, distinguishing embarrassment from other emotions remains an open question. From a behavioral perspective, we tried to ensure that participants reported embarrassment rather than shame by explaining the differences between the two emotions and asking specifically on the VAS scales about how embarrassed they felt at the time. From an acoustic perspective, we used EMO-DB to differentiate between embarrassment and other emotions, but some confounding factors remain, such as dialect differences (EMO-DB used high German vs. our sample was in Swiss German). We assumed that acoustic patterns for embarrassment and reference speech are highly speaker-specific (as is typical for most paralinguistic concepts); hence, the reported UAR rates for the speaker-independent evaluation protocol appeared promising. Given similarities between embarrassment and SA (cf. Hofmann et al.[3]), the current framework is also promising to detect SA in clinical settings. Future studies could apply more advanced speaker-adaptive techniques, incorporating precise acoustic analysis, to further improve classification performance. Lastly, in the light of multidisciplinary approaches, it would also be interesting to compare acoustic indicators with additional (neuro)physiological indicators.

In conclusion, we were able to describe the characteristic way of speaking for people in a state of embarrassment. With a multidisciplinary approach, we have established an effective behavioral paradigm to induce embarrassment and characterize participants' response to the embarrassment induction in acoustic characteristics. Those identified characteristics could be used for the assessment of SA and SAD.

Data availability

The behavioral and machine learning data used to support the findings of this study are available upon request. Contact D.S. for the behavioral data and B.V. for the machine learning data. However, voice data is available only after the approval of the local ethics committee due to the privacy protection of the Human Research Act in Switzerland.

Received: 30 August 2024; Accepted: 11 March 2025
Published online: 20 March 2025

References

1. Miller, R. S. & Tangney, J. P. Differentiating embarrassment and shame. *J. Soc. Clin. Psychol.* **13**, (1994).
2. Tangney, J. P., Miller, R. S., Flicker, L. & Barlow, D. H. Are shame, guilt, and embarrassment distinct emotions? *J. Pers. Soc. Psychol.* **70**, 1256–1269 (1996).
3. Hofmann, S. G., Moscovitch, D. A. & Kim, H. J. Autonomic correlates of social anxiety and embarrassment in shy and non-shy individuals. *Int. J. Psychophysiol.* **61**, 134–142 (2006).
4. Keltner, D. Signs of Appeasement: Evidence for the Distinct Displays of Embarrassment, Amusement, and Shame. in *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)* (eds. Ekman, P. & Rosenberg, E. L.). https://doi.org/10.1093/acprof:oso/9780195179644.003.0007 (Oxford University Press, 2005).
5. Miller, R. S. The interactive origins and outcomes of embarrassment. In: *The Psychological Significance of the Blush* (eds. Crozier, W. R. & Jong, P. J. de) 185–202. https://doi.org/10.1017/CBO9781139012850.013 (Cambridge University Press, 2012).
6. Stocks, E. L., Lishner, D. A., Waits, B. L. & Downum, E. M. I'm embarrassed for you: the effect of valuing and perspective taking on empathic embarrassment and empathic concern. *J. Appl. Soc. Psychol.* **41**, 1–26 (2011).
7. Rozen, N. & Aderka, I. M. Emotions in social anxiety disorder: A review. *J. Anxiety Disord.* **95**, 102696 (2023).
8. Leary, M. R. & Hoyle, R. H. *Handbook of Individual Differences in Social Behavior.* (Guilford, 2013).
9. Miller, R. S. Social anxiousness, shyness, and embarrassability. in Handbook of Individual Differences in Social Behavior 176–191 (The Guilford Press, New York, NY, US, 2009).
10. Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L. & Abramson, A. The voice conveys specific emotions: evidence from vocal burst displays. *Emotion* **9**, 838–846 (2009).
11. Ekman, P. & Friesen, W. V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **17**, 124–129 (1971).
12. Devillers, L. & Vidrascu, L. Real-Life emotion recognition in speech. in Speaker Classification II: Selected Projects (ed Müller, C.) 34–42 (Springer, Berlin, Heidelberg, doi:https://doi.org/10.1007/978-3-540-74122-0_4. (2007).
13. Juslin, P. N. The mirror to our soul?? Comparisons of spontaneous and posed vocal expression of emotion. *J. Nonverbal Behav.* **40** (2018).
14. Patel, S., Scherer, K. R., Björkner, E. & Sundberg, J. Mapping emotions into acoustic space: the role of voice production. *Biol. Psychol.* **87**, 93–98 (2011).
15. Sauter, D. A., Eisner, F., Calder, A. J. & Scott, S. K. Perceptual cues in nonverbal vocal expressions of emotion. *Q. J. Exp. Psychol.* **63**, 2251–2272 (2010).
16. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B. A database of German emotional speech. in *Interspeech 2005* 1517–1520ISCA. https://doi.org/10.21437/Interspeech.2005-446 (2005).
17. Grimm, M., Kroschel, K., Mower, E. & Narayanan, S. Primitives-based evaluation and Estimation of emotions in speech. *Speech Commun.* **49**, 787–800 (2007).
18. Bastin, C., Harrison, B. J., Davey, C. G., Moll, J. & Whittle, S. Feelings of shame, embarrassment and guilt and their neural correlates: A systematic review. *Neurosci. Biobehav Rev.* **71**, 455–471 (2016).
19. Müller-Pinzler, L., Paulus, F. M., Stemmler, G. & Krach, S. Increased autonomic activation in vicarious embarrassment. *Int. J. Psychophysiol.* **86**, 74–82 (2012).
20. Keltner, D., Sauter, D., Tracy, J. & Cowen, A. Emotional expression: advances in basic emotion theory. *J. Nonverbal Behav.* **43**, 133–160 (2019).
21. Weeks, J. W. et al. The sound of fear: assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. *J. Anxiety Disord.* **26**, 811–822 (2012).
22. Kadali, D. B. & Mittal, V. K. Studies on Paralinguistic Sounds, Emotional Speech and Expressive Voices. in *Workshop on Speech, Music and Mind (SMM 2020)* 11–15ISCA. https://doi.org/10.21437/SMM.2020-3 (2020).
23. Mota, N. B., Copelli, M. & Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophr.* **3**, 18 (2017).

24. Mattick, R. P. & Clarke, J. C. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behav. Res. Ther.* **36**, 455–470 (1998).
25. Peters, L., Sunderland, M., Andrews, G., Rapee, R. M. & Mattick, R. P. Development of a short form social interaction anxiety (SIAS) and social phobia scale (SPS) using nonparametric item response theory: the SIAS-6 and the SPS-6. *Psychol. Assess.* **24**, 66–76 (2012).
26. Ouyang, X., Cai, Y. & Tu, D. Psychometric properties of the short forms of the social interaction anxiety scale and the social phobia scale in a Chinese college sample. *Front. Psychol.* **11**, (2020).
27. Delgado, D. A. et al. Validation of digital visual analog scale pain scoring with a traditional Paper-based visual analog scale in adults. *JAAOS Glob Res. Rev.* **2**, e088 (2018).
28. Posit team. *RStudio: Integrated Development Environment for R. Posit Software.* (PBC, 2024).
29. Grimm, M., Kroschel, K. & Narayanan, S. The Vera am Mittag German audio-visual emotional speech database. In *2008 IEEE International Conference on Multimedia and Expo* 865–868. https://doi.org/10.1109/ICME.2008.4607572 (IEEE, Hannover, Germany, 2008).
30. Kehrein, R. Die prosodie authentischer emotionen. *Sprache · Stimme · Gehör*. **27**, 55–61 (2003).
31. Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. in *Proceedings of the 18th ACM international conference on Multimedia* 1459–1462ACM, Firenze Italy, (2010). https://doi.org/10.1145/1873951.1874246
32. Schuller, B. et al. *The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In 148–152.* https://doi.org/10.21437/Interspeech.2013-56 (Lyon, 2013).
33. Yang, S. et al. ISCA,. SUPERB: Speech Processing Universal PERformance Benchmark. in *Interspeech 2021* 1194–1198. https://doi.org/10.21437/Interspeech.2021-1775 (2021).
34. Wagner, J. et al. Dawn of the transformer era in speech emotion recognition: closing the Valence gap. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10745–10759 (2023).
35. Lotfian, R. & Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.* **10**, 471–483 (2019).
36. Chen, S. et al. WavLM: Large-Scale Self-Supervised Pre-Training for full stack speech processing. *IEEE J. Sel. Top. Signal. Process.* **16**, 1505–1518 (2022).
37. Jefferies, P. & Ungar, M. Social anxiety in young people: A prevalence study in seven countries. *PLOS ONE* **15**, e0239133 (2020).
38. Eyben, F. *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. https://doi.org/10.1007/978-3-319-27299-3 (Springer International Publishing, 2016).
39. Hermansky, H. Should recognizers have ears? *Speech Commun.* **25**, 3–27 (1998).
40. Zhang, Y., Zou, J. & Ding, N. Acoustic correlates of the syllabic rhythm of speech: modulation spectrum or local features of the Temporal envelope. *Neurosci. Biobehav Rev.* **147**, 105111 (2023).
41. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
42. G Childers, D., P Skinner, D. & C Kemerait, R. The cepstrum: A guide to processing. *Proc. IEEE.* **65**, 1428–1443 (1977).
43. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **87**, 1738–1752 (1990).
44. Makhoul, J. Linear prediction: A tutorial review. *Proc. IEEE.* **63**, 561–580 (1975).
45. Drahota, A., Costall, A. & Reddy, V. The vocal communication of different kinds of smile. *Speech Commun.* **50**, 278–287 (2008).
46. El Ayadi, M., Kamel, M. S. & Karray, F. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognit.* **44**, 572–587 (2011).
47. Clark, D. M. & Wells, A. A cognitive model of social phobia. in Social Phobia: Diagnosis, Assessment, and Treatment (eds Heimberg, R. G., Liebowitz, M. R., Hope, D. & Schneider, F.) 69–93 (Guilford Press, New York, NY, (1995).
48. Bas-Hoogendam, J. M., van Steenbergen, H., van der Wee, N. J. A. & Westenberg, P. M. Not intended, still embarrassed: social anxiety is related to increased levels of embarrassment in response to unintentional social norm violations. *Eur. Psychiatry*. **52**, 15–21 (2018).
49. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). (2022).
50. Dijk, C., Buwalda, F. M. & de Jong, P. J. Dealing with fear of blushing: A psychoeducational group intervention for fear of blushing. *Clin. Psychol. Psychother.* **19**, 481–487 (2012).

Acknowledgements

We thank Lara Dinner for helping us with the data acquisition. Furthermore, we want to thank all our brave participants, without whom this study would not have been possible.

Author contributions

The conceptualization of this project was done by D.S. and Y.M. The data acquisition for the sample was done by D.S., Y.M., and M.S.The calculations of the behavioral data were done by D.S. with the support of Y.M.The calculations of the machine learning data were done by B.V. with the support of M.MD.The manuscript was written by D.S. with the support of Y.M. and B.V. with the support of M.MD.All authors read and approved the final manuscript.

Funding

UPD university-based fund. B.V and M.MD's work was partially funded through the SNSF Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. 40B2–0_194794).

Declarations

Competing interests

The authors declare no competing interests.

Ethics declarations

The ethics committee of the faculty of human sciences at the University of Bern approved the study protocol (2020-06-00004). All participants gave written informed consent to participate in the study. We hereby confirm that the experiment was performed in accordance with the relevant guidelines and regulations according to the Declaration of Helsinki.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94051-9.

Correspondence and requests for materials should be addressed to D.Š.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.