Annotations from speech and heart rate: impact on multimodal emotion recognition

Sharma, Kaushal; Chanel, Guillaume

How to cite

SHARMA, Kaushal, CHANEL, Guillaume. Annotations from speech and heart rate: impact on multimodal emotion recognition. In: ICMI '23: Proceedings of the 25th International Conference on Multimodal Interaction. Paris. [s.l.] : ACM, 2023. p. 51–59. doi: 10.1145/3577190.3614165

This publication URL: https://archive-ouverte.unige.ch/unige:173677

Publication DOI: 10.1145/3577190.3614165

Annotations from speech and heart rate: impact on multimodal emotion recognition

Kaushal Sharma*
Guillaume Chanel*
0sharmakaushal0@gmail.com
guillaume.chanel@unige.ch
SIMS lab, Computer Science Department, University of Geneva
Carouge, Geneva, Switzerland

ABSTRACT

The focus of multimodal emotion recognition has often been on the analysis of several fusion strategies. However, little attention has been paid to the effect of emotional cues, such as physiological and audio cues, on external annotations used to generate the Ground Truths (GTs). In our study, we analyze this effect by collecting six continuous arousal annotations for three groups of emotional cues: speech only, heartbeat sound only and their combination. Our results indicate significant differences between the three groups of annotations, thus giving three distinct cue-specific GTs. The relevance of these GTs is estimated by training multimodal machine learning models to regress speech, heart rate and their multimodal fusion on arousal. Our analysis shows that a cue(s)-specific GT is better predicted by the corresponding modality(s). In addition, the fusion of several emotional cues for the definition of GTs allows to reach a similar performance for both unimodal models and multimodal fusion. In conclusion, our results indicates that heart rate is an efficient cue for the generation of a physiological GT; and that combining several emotional cues for GTs generation is as important as performing input multimodal fusion for emotion prediction.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

affective computing; machine learning; multimodal fusion; dataset; annotations; social signals; social cues

ACM Reference Format:
Kaushal Sharma and Guillaume Chanel. 2023. Annotations from speech and heart rate: impact on multimodal emotion recognition. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614165

*Both authors contributed equally to this research.

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614165

1 INTRODUCTION

The conscious perception of one's emotional state is pivotal not only for a sense of belonging and well being, but also for rational decision-making [40]. Hence, building computers that can recognize and express affect and consequently interact efficiently with humans can help not only to enhance the emotional experience but also the quality of life.

As addressed by the appraisal theory of emotion [48], emotion is an event-focused, two-step process consisting of: 1) relevance based emotion elicitation mechanisms that 2) share emotional response at the four levels as shown in Figure 1. In their work [42],

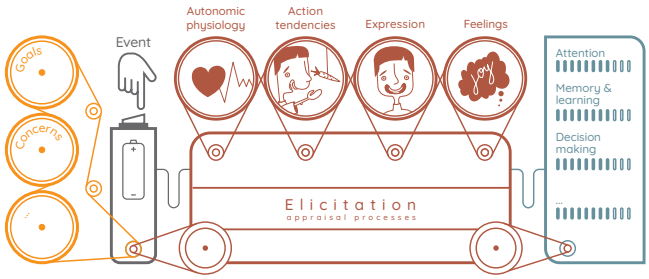

Figure 1: Illustration of the processes involved during the emotion elicitation according to the Appraisal theory of emotion [42] (permission granted by the authors, Eva R Pool and David Sander).

Eva R Pool and David Sander explain how these four levels can be measured. The autonomic level can be measured with peripheral physiological signals such as heart rate and skin conductance. The action tendency can be measured by recording eye movements towards outcome. At an expression level, emotion response can be measured via facial expressions and speech. And finally, the subjective feelings can be estimated via ratings of intensity and pleasantness of the stimuli. By leveraging the multi-level aspect of the emotional response, multi-modal emotion recognition (MER) systems make use of several signals like facial videos, speech, gaze, and physiological activity to make accurate predictions of the subjective feeling. However, to build such systems using supervised machine learning algorithms, we need a large quantity of labelled training data.

The subjective feeling can be measured effectively via three dimensions of arousal, valence, and dominance ratings [2, 47]. Accurate estimates of these ratings are critical as they represent the Ground Truth (GT) for the Machine Learning (ML) based emotion

recognition systems. As a common practice in affective computing, these estimates are collected with the help of external annotators by showing videos of participants' facial expressions and speech. This strategy relies on the natural process of interpreting others' emotions using different social cues such as their facial expressions, speech, hand gestures and body movements [27]. In their study elucidating social signal processing, Marc Mehu et al. [35] differentiate between social signals and social cues. While social signals have evolved explicitly to communicate the state of an underlying process, social cues are subjected to receiver's own interpretation and perception.

Relying on social signals and cues can suffer from several limitations. Firstly, the ratings provided by the annotator only account for the perceived emotion which can be biased by the annotator's background, social environment and emotional state. Moreover, the social cue, in itself can plausibly result in a different interpretation of the underlying emotion. Hence, it may or may not be equivalent to the subject's subjective feeling. Secondly, it is known that while perceiving emotions in a multi-sensory setting, the annotators tend to pay more attention to a particular cue [51]. For instance, if it is assumed that between facial and speech cues, more attention is paid to the speech cue for annotating, then it is plausible that the MER performance with or without the signal corresponding to the more attended cue (in our example, the speech signal) is different, probably higher for speech. This would render the other modalities less effective for MER systems.

This calls for a systematic study of the effect of social signals and cues on the definition of GT. In addition, if several social signals can generate different GTs, there is a need to study how they can be predicted by several modalities and their fusion.

Specifically, the contributions of this paper are as follows:

(1) Estimations of perceived arousal are obtained by using auditory feedback of heart rate as a social cue and its effectiveness for GT estimation is demonstrated.
(2) Arousal annotations are collected from three groups of social cues: speech only, heartbeat sound only, and speech with heartbeat sound, thus producing three GT sets.
(3) The annotations were analyzed to validate that:
   (a) the annotations obtained from the different cues are different from each other.
   (b) inter-rater agreement is similar for the annotations obtained from different social cues.
(4) We run unimodal and multimodal regression models based on Support Vector Regressor (SVR) [1] to predict the three GT sets to estimate the impact of the presented social cues (speech and heartbeat) on the classifiers performance.

2 RELATED WORK

To understand the interplay of various modalities used for MER, there are several studies highlighting various fusion strategies [7]. Yuxuan Zhao et al. [55] make use of EEG, ECG and GSR signals from DEAP [23] and AMIGOS [38] databases to train deep neural networks to classify affective states in the four quadrants formed by the arousal-valance dimensional scale. They demonstrate a superior performance when employing a biologically inspired decision level fusion method for emotion classification. Asif Iqbal Middya et al. [37] use the audio and visual recordings from RAVDESS [31] and SAVEE [19] datasets to train CNN-LSTM based deep networks to classify emotion on a discrete scale. They employ model level fusion in which features learned from the individual modalities separately are thereafter concatenated to classify emotion.

The training of machine learning models for MER necessitate the acquisition of an emotional ground truth. To construct the ground truth, the labels can be predetermined before emotion elicitation e.g., by showing labelled set of images or videos. AMIGOS [38] is one such dataset in which emotions are induced by showing emotional movie excerpts. However, such labels do not encompass the real-world dynamics. To overcome this challenge, many datasets have been developed in a naturalistic and spontaneous emotion inducing settings [50]. To estimate the ground truth for such datasets, two common strategies are used: 1) self-reports in which participants themselves report on the felt emotions during or after the experiment, and 2) external annotations wherein annotators provide their estimation of the perceived emotion using social cues. The latter has more preference for in-the-wild datasets such as RECOLA [45], the OMG-emotion behavior dataset [4], MOSI [54]. Such datasets truly encompass the real-world dynamics as they record videos from social media platforms, movies and reality TV shows. Since they are collected in uncontrolled environment, they are most difficult to predict as well [24]. So far, the social cues shown to the annotators are limited to facial expressions and speech.

Shun Katada et al. [22] employs Multimodal Sentiment Estimation using the multimodal chat-dialog corpus Hazumi1911 [21]. The dataset is annotated by 1) participants themselves and by 2) external annotators using participants' linguistic, speech and facial information to estimate sentiment. Unimodal and multimodal machine learning based models were trained to recognize both self-reported and externally annotated ground-truth using linguistic, physiological, audio and visual features. Results demonstrate that physiological signals were better to predict self-reported annotations while speech performed better on external annotations.

Kosmos et al. [41] explore multimodal interaction for arousal estimation using the RECOLA dataset [45] which contains multimodal recordings in a naturalistic dyadic interaction settings. The dataset is externally annotated using two social signals: participants' speech and facial videos. Results indicate that the machine learning models achieve the highest accuracy for the audio modality. Additionally, when physiological signals are considered, there is no performance gain as compared to the baseline models.

A similar effect can be seen in several studies. For multimodal datasets in which ground truth (GT) is collected via self-reports, performance of physiological signals either is comparable to the other audio/visual modalities or its addition helps to boost the performance [10, 12]. Whereas, it makes no to marginal difference for datasets in which GT is collected via external annotators [39, 43] using participants' facial and audio information. The noteworthy point is that none of the datasets make use of heart rate for to establish GT.

One own's affect state can be experienced by oneself. Moreover, there is strong theoretical evidence for a positive association between interoceptive accuracy and emotion recognition. This is confirmed by empirical findings from psychological studies which used

heartbeat perception task as a measure of interoception accuracy [13, 14, 30, 49] while also revealing that this relationship is complex. Whether inherent interoception capabilities can be extended to infer others' emotions using feedback from their heart rate remains an open and under-researched question. Joris H. Janssen et al. [20] probes this possibility and found that people relate increases in heart rate to increases in emotional intensity.

Considering the findings reported above, we put forth our research questions:

RQ1 - Can heart rate feedback be used as a social cue to annotate arousal?
RQ2 - Do different social cues induce a difference in the annotators' perception of arousal, thus giving different GT sets?
RQ3 - Does the multimodal and unimodal emotion recognition performance differ for those GT sets?

3 METHODS

This section describes the methodology adopted for the experimental protocol, the annotations analysis and the Machine learning models employed to predict arousal.

3.1 Dataset

For the purposes of our study, we use the EATMINT database [8]. It contains multimodal and synchronous recordings of 30 remote dyadic interactions including electrocardiograms (ECG), electrodermal activity, blood volume pulse, respiration, skin temperature, eye-movements, facial expressions, speech signals, French transcripts of the conversation, and screen recordings of the collaborative task. Each dyad had to collaborate for more than 30 minutes to design a slogan against violence at school using the Drew software. Drew [8] is a collaborative environment which allows to build argumentative graphs to discuss ideas and reach a consensus. Since the data is recorded in collaborative dyadic interaction settings, it allows for naturalistic and spontaneous emotion emergence. Although participants in the EATMINT experiment had to report their emotions sporadically during and after the interaction, there are no continuous annotations available for this database.

3.2 Experiment protocol

In order to analyze the effect of several social cues on annotations we collected estimations of arousal levels for the first 10 minutes of EATMINT interactions. We have used only the first 10 minutes to keep our crowdsourcing experiment short and engaging, while also allowing the annotators to follow the interaction from the beginning. More specifically three groups of videos representing the dyadic interaction were created for the annotation purposes. Each group of videos contained common contextual information and specific social cues: speech, heartbeat sound, and a combination of both.

For our study, we make use of speech and ECG signals. While it is a standard practice to use speech for collecting annotations, this study proposes to use heartbeat sounds as social cues. We chose to provide auditory feedback of heart rate to compare two audio cues and avoid cross-modalities effects. We have focused on the estimation of time-continuous perceived arousal as previous studies have shown that both physiological signals and speech can be predictive of arousal [33, 53], while physiological signals are often less efficient to predict valence [6]. Before we can use the signals to construct group specific cues, we apply several preprocessing steps.

Table 1: Properties of S1 and S2 heart Sounds

Sound	Duration	Amplitude Scaling	Frequency band
S1	0.11s	1	10-100 Hz
S2	0.07s	0.6	120-250 Hz

Since we are not interested in studying dyadic interactions, we treat each EATMINT participant independently. Thus for every participant following two strategies are employed to minimize cross-speaker traces. Firstly, speech is cleaned by removing background noise using ffmpeg[1]. We pass the audio through the afftdn filter (nf=-25:nr=50). The parameters are chosen empirically by visual inspection of the signals. This first filtering is followed by a band-pass filter with cutoff frequencies between 200 Hz and 3000 Hz. Secondly, using exact timestamps from transcripts, the speech signal is silenced for the time duration when the other speaker is talking. All the speech signals are then normalized for loudness using Audacity's Loudness normalization effect with default settings. The signals are sampled at 22050 Hz .

To give the participants a natural feel of the heartbeat sound, we decided to synthesize the S1 and S2 sounds from the ECG signal. The S1 and S2 sounds are the main components of the heart sounds [34] and are colloquially referred to as the "lub-dub" heart sound. Technically, S1 occurs when the tricuspid and mitral valves close marking the beginning of systole and S2 occurs when aortic and pulmonary valves close thus marking the end of systole and beginning of diastole. As per the literature [11, 15, 34], the S1 is longer, louder, and occurring in a lower frequency band than S2. The parameter selections for synthesizing S1 and S2 sounds are detailed in Table 1, drawing from the literature. The time duration between the sounds S1 and S2 is taken to be 0.287 secs based on the findings of [34].

As a first step, heart rate is computed from the ECG Signals. This is achieved by performing RR peaks detection employing the algorithm proposed by P. Hamilton [18] using BioSPPy[2] toolbox. The R-peaks are further corrected using median filtering algorithm based on two parameters, *steps* and *threshold*. For a current Inter-Beat-interval (IBI), median value of the *steps* number of IBIs preceding and following the current IBI is calculated. If the current IBI exceeds *threshold* times the calculated median IBI, then equally spaced IBIs are inserted. The duration of the newly inserted IBIs is kept same as the median IBI. Care is taken to keep the shortest IBI above 2 ms. The values of the *steps* and *threshold* used are 4 and 1.7, respectively, which were empirically determined through visual inspection of the R-peaks.

These R-peaks are then converted to wav format following the procedure described below and illustrated in Figure 2.

(1) We start with synthesising white noise by drawing random samples from a normal distribution.

[1]https://ffmpeg.org/
[2]https://pypi.org/project/biosppy/

(2) The white noise is then filtered separately for S1 and S2 based on the respective characteristics of Table 1.
(3) The filtered white noise is passed through the hanning window.
(4) As the original sampling rate of ECG signal is 512, the sampling rate of final wav sound is kept at 512*17 = 8074 Hz. At this sampling rate, we construct a dirac like time-series having unitary values for indices exactly where an R-peak occurred for S1. For S2, the time-series is constructed using shifted R-peaks by 2500 samples or 0.28 s to account for the time duration between S1 and S2 sounds.
(5) The respective dirac like time-series is convolved with the respective windowed and filtered white noise for S1 and S2.
(6) Finally, the convolved S1 and S2 signals are concatenated on the time-axis.
(7) Using the signal above, heart rate sounds are recorded in a mono wav file with sampling rate 8704 Hz.

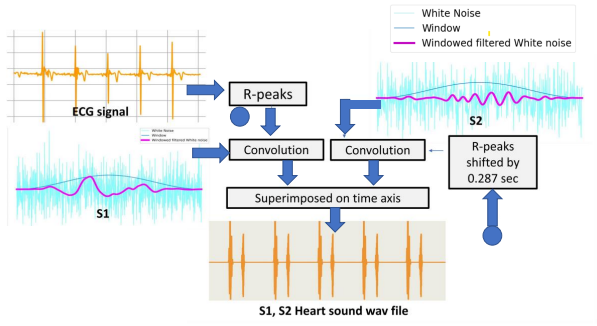

Figure 2: Steps for heart rate sonification

Using the two speech and heartbeat sound cues derived from the participant's speech and ECG signals, we construct the following three groups of social cues: speech only, heartbeat sound only, both speech and heartbeat sound. The speech and heartbeat sound is combined using ffmpeg amerge filter with a stereo output balanced in both ears.

As mentioned above, we treat each EATMINT participant independently. Consequently, the speech and heartbeats of the participant's partner are not added in the stimuli to ensure that the annotations differences are only due to the social cues of the studied participant. Nevertheless, subtitles of the whole conversation (i.e. transcripts of both dyad members) are provided to the annotators in all conditions. In addition, a video of the participant's screen is included in the stimuli so that annotators can obtain information about the arguments and actions of the dyad. Adding subtitles and a video of the screen are considered to be contextual information necessary for the understanding of the dyadic interaction.

Finally, for each participant and group, we combine the group specific cues (speech, heartbeat sound or both) and contextual information in a video representing the first 10 minutes of the interaction. Dummy screens representing the three conditions of the experiment are shown in Figure 3. To overcome the annotator specific bias, each video is annotated by 6 different annotators for each condition. Thus, in total for each EATMINT participant, we collect 18 annotations (3 conditions * 6 annotators). An annotator is neither repeated across videos nor conditions. Because of the large number of annotations, we used two crowd-sourcing platforms: Prolific[3] and Toloka[4] to invite the annotators for our study. The only participant selection criterion used is fluency in French language to understand the conversations of the EATMINT participants. From the 450 annotators data (25 * 6 * 3), we can summarize demographic statistics of 439 annotators as some of the annotator revoked their consent to share demographic information. The age of the participants ranges from 19 to 69 (median age 27). For all the three conditions, the interquartile range for age (heartbeat sound: 10, speech: 8, combined heartbeat sound and speech: 11) is similar. Participants are recruited from 51 different countries with France being the most frequent, accounting for 31 percent.

To obtain time-continuous annotation for arousal, we used PAGAN [36] which is an online platform for collecting continuous dimensional affect annotations. Specifically, we used RankTrace interface (Figure 4) provided by PAGAN as this is shown to induce better inter-rater agreement than the GTrace and BTrace [36]. Before the main task is started, thorough knowledge of arousal and clear instructions are provided. This study has received approval from the ethical committee of the University of Geneva, and participants gave their consent for participation in the study and the use of the recorded data.

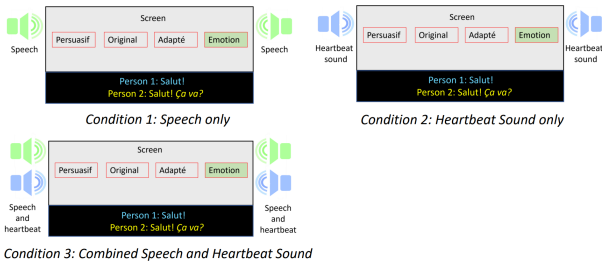

Figure 3: Representations of the screens for the three conditions of the experiment.

Figure 4: Screenshot of the RankTrace interface used for continuous annotations in the PAGAN [36] platform

3.3 Annotation analysis

3.3.1 Post processing of annotations. To synchronize the unevenly sampled annotations provided by different annotators, annotations are interpolated at a frequency of 100 Hz by using a forward fill

[3]https://www.prolific.co/
[4]https://toloka.ai/

method: the value at a given timestamp in the re-sampled time-series is equal to the value of the closest previous timestamp in the original series. To account for variation in human annotators, the data from each annotator is re-scaled in the range [0, 1] using min-max normalization.

3.3.2 Inter-rater agreement. To answer our RQ2 we want to analyse whether there are differences between the annotations obtained from the different social cues. Firstly, we need to define an index that can be used to quantify the differences. Within the framework of the studies, inter-rater agreement (IRA) is defined as the extent to which ratings are identical [16]. Many indices are usable such as Krippendorff's alpha [25], the Intraclass Correlation Coefficient (ICC) [5] and the Concordance Correlation Coefficient (CCC) [28] to measure IRA. A lower value of such an index would indicate lesser agreement and hence more differences. Concordance correlation coefficient (CCC) is a statistical measure which can be used to assess the agreement between the two sets of continuous ratings made by different raters. By measuring the variation from the 45 degree through the origin, it accounts for both the the precision (degree of variability) and the accuracy (scale or location shift) [3, 28]. Lawrence et al. [28] also provides a modified version of CCC which can be used for more than two variables. We have further adapted the formula as advised in [29]. The Concordance Correlation Coefficient, σ_c for P annotators is given by:

$$\sigma_c = \frac{2\sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\sigma_{ij}}{\sum_{i=1}^{P-1}\sum_{j=i+1}^{P}\left(\mu_i-\mu_j\right)^2+(P-1)\sum_{i=1}^{P}\sigma_i^2} \quad (1)$$

where σ_{ij} is the covariance between the annotations of participants i and j, μ_i and σ_i are the mean and variance of participant i , respectively. This form gets reduced to CCC for two variables when we substitute $P = 2$ in equation 1. We have used the biased estimate of CCC.

CCC is preferred over other indices mentioned above as for ICC ANOVA assumptions need to be met whereas CCC provides more usage freedom as it is defined with less assumptions [9]. Moreover, ICC is closely related to CCC [9]. Krippendorff's alpha [26] can be used, however, its estimation is based on the units of analysis which can explode quickly for a continuous variable. Furthermore, its estimations require pairwise calculations to obtain coincidences within units thus making it computationally difficult.

Secondly, to qualify the differences between the annotations for the three conditions, we need a reference level to which we can compare the overall between conditions agreement score and draw relevant conclusions. Hence, we are also interested in calculating agreement among annotators within each condition. If for each condition, the agreement among the annotators within that condition is higher than the overall between conditions agreement, then it can be said that the annotations for the three conditions are different from each other.

Thirdly, to compute CCC, we need a systematic methodology that allows for a fair comparison of the between and within conditions agreement. As will be explained in section 3.4 below, for each condition we have averaged annotations across the six annotators to obtain an unique time-series which is representative of ground truth for that condition. Since the Machine learning analysis is run on these averaged estimates, it is important that we assess the differences among these averaged annotations. Moreover, averaging allows for de-noising the signals and reducing annotators bias. To compute the overall between conditions agreement, one can simply compute CCC for the three averaged time-series, one per each condition, using Equation 1. Similarly, to compute the agreement within the annotators of a given condition, the formula given in Equation 1 can be used directly to compute CCC. One potential limitation of this approach is that it can overestimate agreement between conditions because it incorporates the averaging effect, which reduces the annotators' variability. Consequently, we came up with a procedure which maintains this variability in a comparable manner by allowing for averaging equal number of annotators while computing agreements at the two levels. The methodology is described below.

To compute inter-rater agreement between conditions, for each of the three conditions, we first sample three different annotators from from the six annotators belonging to that condition. Thereafter, the annotations are averaged across the three annotators within a condition to obtain an unique time-series per condition. Lastly, using Equation 1, CCC is computed for the three time-series. This procedure is repeated for all possible combinations of selecting three different annotators from three groups of 6 annotators each to accommodate for the effect of all the annotators. As a final step, the inter-rater agreement score is averaged across all the combinations.

To compute inter-rater agreement among the annotators within a given condition, we start by sampling three different annotators from the 6 annotators belonging to that condition thus generating two sub groups of 3 different annotators each . For each sub-group, annotations are averaged across the three annotators to obtain a unique time-series per sub-group. Lastly, using Equation 1, CCC is computed for the two time-series. This procedure is repeated for all possible combinations of choosing 3 annotators from the six annotators belonging to a given condition. As a final step, the inter-rater agreement score is averaged across all the combinations.

Finally, to assess if agreement is significantly different for within and between levels, we performed paired t-test independently for each condition.

To answer our RQ1, we analyze if the inter-rater agreement score is different for the three groups of annotations. Here, we want to compare the agreement scores obtained for each of the three conditions. To do so, we compute within group inter-rater agreement score as described above. We then performed one-way repeated measures ANOVA to determine whether the mean inter-rater agreement scores are significantly different between the three groups of annotations. If the agreement score is similar for all the three social cues, then it can mean that the annotators were able to interpret each of the social cues similarly.

3.4 Machine learning (ML) analysis

This section describes the methodology adopted to study the third research question RQ3. Since our focus is on comparing performances, we use a baseline ML algorithm: Support Vector Regressor (SVR) [1] to predict perceived arousal. Literature review reveals SVR as a reliable choice for continuous arousal recognition [44]. As we ran 9 models for each of the 25 participants as explained below, SVR

being memory efficient converged faster and saved long training time typical of deep architectures. We use the Radial Basis Function kernel and a unitary value for the regularization parameter. The following models are trained using features from three groups of modalities to predict each of the three GT sets individually: 1) a unimodal model with speech only, 2) a unimodal model with heart rate only, and 3) a multimodal model with speech and heart rate. Thus, in total we run 9 models for each leave-one participant out iteration. The performance of the models is measured by computing Concordance Correlation Coefficient (using equation 1) between the ground truth and predictions.

3.4.1 Feature extraction. We compute hand-crafted features for both modalities: speech and heart rate. We perform leave-one participant out cross validation to generate training and test sets. Both signals are segmented using 10 sec overlapping windows with 20 percent overlap for the train set and no overlap for the test set.

The speech signals are down sampled to 16 kHz to reduce processing times. To minimize speaker level biases, signals are min-max normalized in the range from -1 to 1 before segmentation. Using librosa[5], the following features are extracted for segmented windows: 12 mfcc coefficients, 20 chroma-STFT coefficients, spectral-rolloff, spectral-bandwidth, 7 spectral-centroid coefficients, spectral-contrast, spectral-flatness, zcr and rms. The settings for FFT window size, n-fft are kept as 256 and hop length as 128. Finally, statistical moments (mean, argmax, argmin, maximum, minimum, range, standard deviation, skew, kurtosis, first quartile, second quartile and third quartile) are computed across time frames of a segmented window giving in total 540 features per time window (45 features * 12 statistical moments).

Using the same method as described in section 3.2, the R-peaks of the 512 Hz ECG signal are extracted for each segmented windows. We then perform time-domain analysis of the heart rate variability using neurokit [32], thus giving 23 features derived from the NN intervals. To train multimodal model utilising features from both speech and heart rate, we simply concatenate the features from the individual modalities thus giving 563 features (540 from speech modality and 23 from the heart rate.)

3.4.2 Annotation merge Strategy. The post processing steps as described in section 3.3.1 are applied to the annotations. Furthermore, to compute one GT sequence per participant and per social cue, it is required to merge the annotations of the six annotators. This is computed by using the Evaluator Weighted Estimator (EWE) since it has been shown that it produces more reliable ground truths than computing the mean by maximum likelihood estimation (MLE) [17]. With EWE the combined estimated arousal is weighted by a confidence score computed for each annotator. The confidence score is computed as the correlation between the average annotations and the annotator specific annotations.

After merging the annotations a linear trend was observed in the resulting arousal time series. Since we are interested in capturing short-term emotional dynamics, linear detrending is applied using python library scipy[6]. The detrended time-series is once again scaled between the range 0 to 1 using min-max normalisation. Finally, the annotations are smoothed by applying rolling mean using a window size 10 seconds to match with the signal segmentation window length. The overlap settings are kept same as that for signal segmentation. Using the methods described above, we prepare the data for training.

Finally, we perform two-way repeated measures ANOVA to assess the effect of social cues used for constructing GTs and the training modality used for extracting features on the emotion recognition performance. Appropriate post hoc tests are carried out. Please note that all the statistical tests are checked for underlying assumptions of normality, the absence of extreme outliers, and sphericity. The significance level is kept at 0.05 and is adjusted accordingly when required. Unless stated otherwise, two-tailed tests are performed. We also fit ANOVA model for a within subjects design using the afex[7] package in R [46]. Using the model estimated marginal means we plot an interaction plot with 95% confidence intervals.

4 RESULTS AND DISCUSSIONS

4.1 Annotation analysis results

4.1.1 Comparison of within conditions and between conditions agreements: Using the methodology detailed in section 3.3, we compute the agreements at the two levels: within and between conditions. The results of the paired t-test are shown in Figure 5. It can be seen that for the heartbeat sound only condition, there was a significant difference in mean agreement between the annotators within the condition (M = 0.39, SD = 0.20) and the annotators across all the conditions (M = 0.28, SD = 0.15); t(24) = -3.3, p = 0.003. For the speech only condition, there was a significant difference in mean agreement between the annotators within the condition (M = 0.38, SD = 0.18) and the annotators across all conditions (M = 0.28, SD = 0.15); t(24) = -2.69, p = 0.013. Similarly, for the combined speech and heartbeat sound condition, there was a significant difference in mean agreement between the annotators within the condition (M = 0.36, SD = 0.21) and the annotators across all conditions (M = 0.28, SD = 0.15); t(24) = -2.6, p = 0.016. To summarise, for all the social cues, on average the agreement among the annotators within a condition is significantly higher than that for annotators between conditions. In essence, the variations in annotations obtained under the same condition are comparatively smaller than those gathered under different conditions. This demonstrates that annotations originating from different social cues manifest notable dissimilarities. This affirmative response to RQ2 confirms that different social cues induce different perceptions of arousal among annotators. This in turn support the creation of distinct ground truth sets each holding information specific to the social cue.

4.1.2 Comparison of within conditions agreements: Using methods described in section 3.3, we compute within condition agreements. As illustrated in Figure 5, the mean inter-rater agreement scores are 0.39 for heartbeat sound only, 0.38 for speech only, and 0.36 for the combined speech and heartbeat sound conditions. Prior values for IRA using CCC are ranging between 0.277 and 0.431 [52]. The values obtained here are within this range. We perform a one-way repeated measures ANOVA test to compare the within-group inter-rater

[5]https://librosa.org/
[6]https://scipy.org/
[7]https://cran.r-project.org/web/packages/afex/index.html

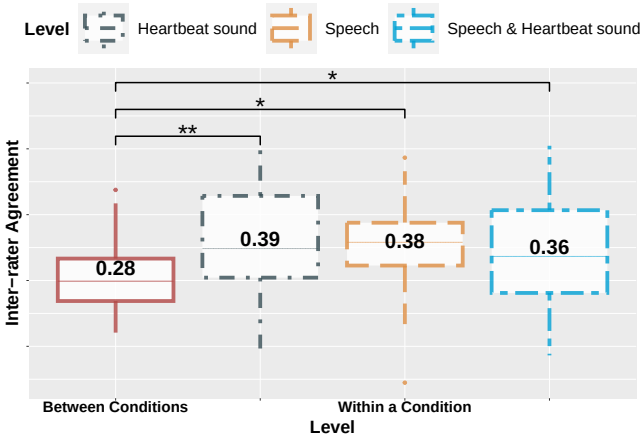

Figure 5: Comparison of within and between group inter-rater agreement scores for the annotations collected from the three groups of social cues.

agreement scores for the annotations from the three social cues. The statistical analysis reveals that there is no significant difference in agreement among annotations obtained from different social cues, $F(2,48) = 0.34$, $p = 0.72$. Consequently, there is insufficient statistical evidence to reject the null hypothesis, suggesting equality in group means. This result supports the hypothesis that annotators were equally able to interpret both heartbeat sound and speech cues. This finding partially affirms the response to RQ1, demonstrating the viability of utilizing heart rate feedback as a social cue for annotating arousal.

4.2 ML analysis results

Using methods described in section 3.4, a two-way repeated measures ANOVA was performed to assess the effect of social cues used for constructing GTs and the training modality used for extracting features on the emotion recognition performance. To facilitate discussion, we use following acronyms for the three groups of modalities: 1) *sp modality* - a unimodal model with speech only, 2) *hr modality* - a unimodal model with heart rate only, and 3) *sp+hr modality* - multimodal model with speech and heart rate. Results of the ANOVA test show that there was a statistically significant two-way interaction between GT social cue and the training modality on performance, $F(2.51, 60.22) = 26.91$, $p < .001$. The Greenhouse–Geisser method was applied for Sphericity correction. The interaction plot using the model estimated marginal means and the 95% confidence intervals is shown in Figure 6.

As part of the post hoc tests, the effect of training modality was analysed for each of the GT social cue. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of modality was significant for the speech cue ($p < 0.0001$) and heartbeat sound cue ($p < 0.001$) but not for the combined speech and heartbeat sound cue ($p = 0.174$). Hence, pairwise comparisons, using paired t-test, were performed for the speech only and heartbeat sound only cues.

We first analyse the results of the pair-wise comparisons for the heartbeat sound only cue. Results show that the performance was significantly different between 1) *sp modality* and *hr modality* ($p < 0.001$), and 2) *sp modality* and *sp+hr modality* ($p < 0.00001$). The performance was not significantly different between *hr modality* and *sp+hr modality* ($p > 0.3$). To sum up, to predict arousal estimated from the heartbeat sound cue, there is a significant difference in performances of the two classifiers: 1) trained using features from the corresponding modality: heart rate, and 2) trained using features from modalities other than heart rate. At the same time, if both the classifiers are trained using features from the corresponding modality, then then the comparison is insignificant.

We now analyse the results of the pair-wise comparisons for the Speech only cue. The performance is significantly different for 1) *hr modality* and *sp modality* ($p < 0.01$), and 2) *hr modality* and *sp+hr modality* ($p < 0.001$). The performance was found to be not significantly different between *sp modality* and *sp+hr modality* ($p > 0.1$). We observe the same effect as that for heartbeat sound only cue. To predict arousal estimated from the speech only cue, classifiers trained using features from the corresponding modality:speech perform better in comparison to those which haven't used the features from speech.

To predict arousal estimated from combined speech and heartbeat sound cue, there was no statistically significant difference in performance of all the three modalities: *sp modality*, *hr modality* and *sp+hr modality*. This is also in-line with the corresponding modality effect. Since the cue has been estimated using both the speech and the heart rate, all the three modalities perform similarly.

Collectively, the findings suggest that specific ground truth sets are more accurately predicted when corresponding modalities are utilized. This explains the low performance of physiological signals for predicting emotions which were externally annotated using facial and speech cues [39, 41, 43]. In their works, adding of physiological signals in a multimodal paradigm did not improve performance because physiological information was not present in the GT. The earlier discussion has highlighted that different social cues influence annotators' perception of arousal, consequently yielding distinct GT sets. Our results provide further evidence to validate RQ3, emphasizing that the performance difference between multimodal and unimodal emotion recognition is subjected to the GT employed. Fusion of multiple modalities to enhance performance is more effective if the GT has also been estimated using the corresponding social cues from the modalities.

In section 4.1.2 we discussed the viability of employing heart rate feedback as a social cue for annotating arousal. Through our analysis of emotion recognition performance, we observe that for combined speech and heartbeat sound cue, the performance of the speech modality, heart rate modality and their multimodal fusion exhibit no significant differences. This indicates that annotators effectively integrated information from both cues to form a coherent GT, yielding comparable predictions from both unimodal and multimodal emotion recognition systems. This positive outcome directly addresses RQ1, substantiating that heart rate feedback is not only as viable as speech cues, but indeed beneficial for emotion recognition, contributing valuable information to the ground truth. We possess unique inherent ability to sense our well-being via interoception of which heart rate feedback is an integral part [13]. It seems that it is possible to extend that capability to infer others emotion via feedback of their cardiac activity. In addition, it

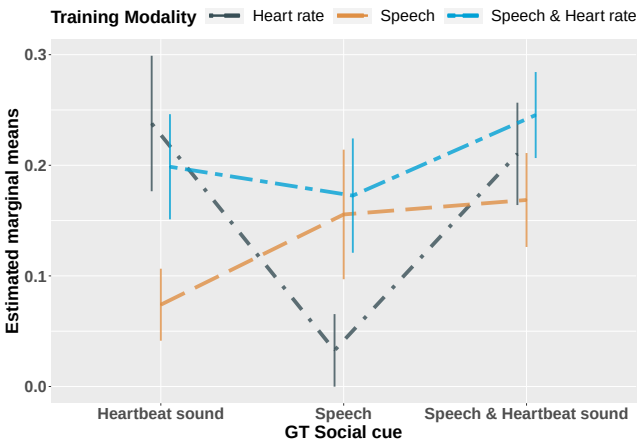

Figure 6: Interaction plot for the effect of training modality and the GT social cue on emotion recognition performance.

is possible that this competence improves with the widespread use of wearable sensor technology in day to day lives.

4.3 Conclusions

In conclusion, we performed a systematic analysis of the effect of social cues on ground-truth definition and on the performance of emotion recognition systems. A subset of the EATMINT database was continuously and externally annotated for arousal levels. The annotations were collected from three groups of social cues: 1) heartbeat sound only, 2) speech only, and 3) combined speech and heartbeat sound. Unimodal and multimodal machine learning models were trained using features from speech and heart rate to predict these three GT sets.

Our results show that the annotations collected for different social cues are significantly different from each other. Specifically, for all three social cues, the mean agreement among annotators within a condition was found to be significantly higher than the agreement among annotators across all conditions. This suggests that the perception of the same underlying emotion varies depending on the social cues considered. While an individual might experience a distinct emotion internally, its external estimation varies according to the employed social cue. Therefore, while establishing ground truth standards for emotion recognition, it is crucial to account for the influence of different social cues on emotion perception. These insights contribute to a deeper understanding of how various cues impact emotion interpretation. Furthermore, our analysis show that the annotators reach a similar level of agreement for heartbeat sound and speech cues. Hence using auditory feedback of heart-rate is as feasible as the conventionally used speech cue.

The analysis of machine learning performance in our study highlights the superior predictive capability of modality-matched ground truth in our study. This establishes the relevance of the differences observed between the cue-specific ground truth sets from an emotional recognition perspective. While much attention has been directed towards the multimodal aspects of the emotion recognition, our findings emphasize that the inclusion of multiple modalities yields greater benefits when the ground truth also incorporates cues from respective modalities. Moreover, our study underscores the viability and utility of utilizing auditory feedback of heart rate for emotion recognition. It explains why the performance of multimodal emotion recognition using physiological signals makes marginal difference to predict emotions which have been externally annotated using only facial and speech cues. A recommendation of this paper is thus that physiological feedback should be part of the social cues given to annotators in the context of external emotion annotation tasks.

For our study we could not use facial expression because of ethical reasons, but it would be interesting to conduct a study to see if similar results can be obtained. Moreover, analysing if there are social cues which are more efficient for a given dimension of emotion (e.g. facial expressions for valence and physiological activity for arousal) can help towards building robust affective computers. While we used baseline models to perform our analysis, it would be interesting to see if the effects studied here are modulated by the complexity and sophistication of the architecture used to recognize emotion. Our analysis shows that unimodal cue specific GTs are relevant for emotion recognition. Similar to how several modalities are fused together to construct a unified learned representation, deep-architectures can be employed to learn to predict the several cue-specific GTs in a multitask learning paradigm. This might have advantages of learning robust representations from multimodal signals and cues. The findings presented in this study opens new ways to advance in the field of affective computing and consequently Human Computer Interaction.

ACKNOWLEDGMENTS

This work is co-financed by Innosuisse, project 34316.1 IP.ICT.

REFERENCES

[1] Mariette Awad and Rahul Khanna. 2015. *Support Vector Regression.* Apress, Berkeley, CA, 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4
[2] Iris Bakker, Theo Van Der Voordt, Peter Vink, and Jan De Boon. 2014. Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology* 33 (2014), 405–421.
[3] Huiman X Barnhart, Michael Haber, and Jingli Song. 2002. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 58, 4 (2002), 1020–1027.
[4] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The OMG-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN).* IEEE, 1–7.
[5] John J Bartko. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological reports* 19, 1 (1966), 3–11.
[6] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S. Huang. 2016. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (Amsterdam, The Netherlands) *(AVEC '16).* Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/2988257.2988264
[7] Yujian Cai, Xingguang Li, and Jinsong Li. 2023. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors* 23, 5 (2023), 2455.
[8] Guillaume Chanel, Mireille Bétrancourt, Thierry Pun, Donato Cereghetti, and Gaëlle Molinari. 2013. Assessment of computer-supported collaborative processes using interpersonal physiological and eye-movement coupling. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* IEEE, 116–122.
[9] Chia-Cheng Chen and Huiman X. Barnhart. 2008. Comparison of ICC and CCC for assessing agreement for data without and with replications. *Computational Statistics & Data Analysis* 53, 2 (2008), 554–564. https://doi.org/10.1016/j.csda.2008.09.026

[10] Joaquim Comas, Decky Aspandi, and Xavier Binefa. 2020. End-to-end facial and physiological model for affective computing and applications. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 93–100.

[11] S.M. Debbal and F. Bereksi-Reguig. 2007. Time-frequency analysis of the first and the second heartbeat sounds. *Appl. Math. Comput.* 184, 2 (2007), 1041–1052. https://doi.org/10.1016/j.amc.2006.07.005

[12] Vipula Dissanayake, Sachith Seneviratne, Hussel Suriyaarachchi, Elliott Wen, and Suranga Nanayakkara. 2022. Self-supervised Representation Fusion for Speech and Wearable Based Emotion Recognition. *Proc. Interspeech 2022* (2022), 3598–3602.

[13] Barnaby D Dunn, Hannah C Galton, Ruth Morgan, Davy Evans, Clare Oliver, Marcel Meyer, Rhodri Cusack, Andrew D Lawrence, and Tim Dalgleish. 2010. Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological science* 21, 12 (2010), 1835–1844.

[14] Wendy D'Andrea, Nadia Nieves, and Treva Van Cleave. 2022. To thine own self be true: interoceptive accuracy and interpersonal problems. *Borderline personality disorder and emotion dysregulation* 9, 1 (2022), 1–7.

[15] Asir B El, L Khadra, AH Al-Abbasi, et al. 1996. Time-frequency Analysis of Heart Sounds. (1996).

[16] Natasa Gisev, J. Simon Bell, and Timothy F. Chen. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9, 3 (2013), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

[17] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 381–385.

[18] P. Hamilton. 2002. Open source ECG analysis. In *Computers in Cardiology*. 101–104. https://doi.org/10.1109/CIC.2002.1166717

[19] Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).

[20] Joris H Janssen, Wijnand A Ijsselsteijn, Joyce HDM Westerink, Paul Tacken, and Gert-Jan de Vries. 2013. The tell-tale heart: perceived emotional intensity of heartbeats. *International Journal of Synthetic Emotions (IJSE)* 4, 1 (2013), 65–91.

[21] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 315–323.

[22] Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation. *IEEE Transactions on Affective Computing* (2022).

[23] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[24] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* 127, 6-7 (2019), 907–929.

[25] Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement* 30, 1 (1970), 61–70.

[26] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[27] Jens Lange, Marc W. Heerdink, and Gerben A. van Kleef. 2022. Reading emotions, reading people: Emotion perception and inferences drawn from perceived emotions. *Current Opinion in Psychology* 43 (2022), 85–90. https://doi.org/10.1016/j.copsyc.2021.06.008

[28] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. http://www.jstor.org/stable/2532051

[29] Lawrence I-Kuei Lin. 2000. A note on the concordance correlation coefficient. *Biometrics* 56, 1 (2000), 324–325. https://doi.org/10.1111/j.0006-341X.2000.00324.x

[30] Alexander Lischke, Rike Pahnke, Anett Mau-Moeller, Robert Jacksteit, and Matthias Weippert. 2020. Sex-specific relationships between interoceptive accuracy and emotion regulation. *Frontiers in Behavioral Neuroscience* 14 (2020), 67.

[31] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[32] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[33] Mizuki Matsubara, Olivier Augereau, Charles Lima Sanches, and Koichi Kise. 2016. Emotional arousal estimation while reading comics based on physiological signal analysis. In *Proceedings of the 1st International Workshop on comics ANalysis, Processing and Understanding*. 1–4.

[34] Steven McGee. 2018. Chapter 40 - The First and Second Heart Sounds. In *Evidence-Based Physical Diagnosis (Fourth Edition)* (fourth edition ed.), Steven McGee (Ed.). Elsevier, Philadelphia, 333–344.e2. https://doi.org/10.1016/B978-0-323-39276-1.00040-8

[35] Marc Mehu and Klaus R Scherer. 2012. A psycho-ethological approach to social signal processing. *Cognitive processing* 13 (2012), 397–414.

[36] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video affect annotation made easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 130–136.

[37] Asif Iqbal Middya, Baibhav Nag, and Sarbani Roy. 2022. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems* 244 (2022), 108580.

[38] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* 12, 2 (2018), 479–493.

[39] Sabrina Patania, Alessandro D'Amelio, and Raffaella Lanzarotti. 2022. Exploring Fusion Strategies in Deep Multimodal Affect Prediction. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*. Springer, 730–741.

[40] R.W. Picard. 1995. *Affective Computing*. MIT Media Laboratory Perceptual Computing Section Technical Report 321. MIT Boston, MA.

[41] Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2022. Supervised contrastive learning for affect modelling. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 531–539.

[42] Eva R Pool and David Sander. 2021. Emotional learning: measuring how affective values are acquired and updated. In *Emotion Measurement*. Elsevier, 133–165.

[43] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (2015), 22–30.

[44] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*. 3–9.

[45] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–8. https://doi.org/10.1109/FG.2013.6553805

[46] Team RStudio et al. 2020. RStudio: integrated development for R. *Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com* (2020).

[47] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[48] Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

[49] Stephanie A Schuette, Nancy L Zucker, and Moria J Smoski. 2021. Do interoceptive accuracy and interoceptive sensibility predict emotion regulation? *Psychological Research* 85 (2021), 1894–1908.

[50] Mohammad Faridul Haque Siddiqui, Parashar Dhakal, Xiaoli Yang, and Ahmad Y Javaid. 2022. A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. Multimodal Technologies and Interaction 6, 6 (June 2022), 47.

[51] Sachiko Takagi, Saori Hiramatsu, Ken Ichi Tabei, and Akihiro Tanaka. 2015. Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. *Frontiers in Integrative Neuroscience* 9 (2 2015). Issue FEB. https://doi.org/10.3389/fnint.2015.00001

[52] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.

[53] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[54] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[55] Yuxuan Zhao, Xinyan Cao, Jinlong Lin, Dunshan Yu, and Xixin Cao. 2021. Multimodal affective states recognition based on multiscale cnns and biologically inspired decision fusion model. *IEEE Transactions on Affective Computing* (2021).