# FUSE - FrUstration and Surprise Expressions: A Subtle Emotional Multimodal Language Corpus

**Rajesh Titung, Cecilia O. Alm**
Rochester Institute of Technology
New York, USA
{rt7331, coagla}@rit.edu

**Abstract**

This study introduces a novel multimodal corpus for expressive task-based spoken language and dialogue, focused on language use under frustration and surprise, elicited from three tasks motivated by prior research and collected in an IRB-approved experiment. The resource is unique both because these are understudied affect states for emotion modeling in language, and also because it provides both individual and dyadic multimodally grounded language. The study includes a detailed analysis of annotations and performance results for multimodal emotion inference in language use.

**Keywords:** corpus, multimodal affective computing, subtle emotion, frustration, surprise

## 1. Introduction

Human emotional expressions are often subtle in language interactions (Yitzhak et al., 2017). They are influenced by many factors, such as sociocultural norms, interpersonal roles, and the situation in which the discourse occurs (Keltner et al., 2019; Parkinson et al., 2005), and they involve ambiguity in expression or interpretation (Alm, 2011). This poses challenges for eliciting emotional language and also for modeling non-acted emotional language, which contrasts with exaggerated expressions obtained from controlled conditions such as scripted or improvisational acted dialogues (Zeng et al., 2009), whose characteristics do not generalize to nuanced, real-world language. In addition, there is a need for emotional language corpora capturing non-basic emotions and non-acted dialogue (Alm, 2022). Addressing this gap, we introduce a new multimodal corpus for expressive task-based spoken language and dialogue, focused on *frustration* and *surprise* reactions – two understudied emotions which are of particular importance for human-computer teaming and interaction, yet challenging to elicit or characterize. For instance, in elicitation, it is usually not possible to be surprised more than once by repeated stimuli (Niepel et al., 1994; Shea et al., 2018), and frustration may combine with other emotions having incongruent polarities, such as with amusement, perhaps to release tension arising from frustration (Morreall, 1982).

This study[1] is valuable because surprise and frustration are understudied for emotion modeling in language, and because it focuses on both individual and dyadic interactions and language that is multimodally grounded.[2] The tasks were also designed to elicit low-arousal affect states in speakers. For example, motivated by prior literature, continuously blocking speakers to achieve their task goals helps elicit accumulated frustration (Amsel, 1992), while unexpected stimuli elicit nuanced surprise (Meyer et al., 1997). This was an IRB-approved study with informed consent.

We focus on natural language, yet ensure that it is multimodally grounded since a single modality can be insufficient and, at times, even misleading (Hoque and Picard, 2011). The combination of multiple modalities can also help improve the robustness of systems that seek to respond to affective behaviors (Liu et al., 2016). D'Mello and Kory (2015) reported that the use of multimodal data allowed their models to reliably achieve improved results, compared to unimodal models.

Prior multimodal affective corpora tend to lack spontaneity and be collected under acted, controlled conditions, resulting in premeditated emotional responses (Busso et al., 2008). Models trained on such data may appear high-performing but may fail to generalize to speakers' subtler expressions. On the other hand, models using emotional language from the wild (e.g., distressed calls) raise ethical concerns. Our study strikes the balance between realistic spontaneity in lab-based tasks drawing on a social context, resulting in high ecological validity (Kory and D'Mello, 2014).

Finally, it is questionable to annotate, or guess, speakers' emotions from post-inspecting their data. Thus, for emotional state annotation, traditional annotation with independent raters often fails. Also, subjective preferences or cultural differences among annotators versus those who produced the language can result in misinterpretation (Barrett,

[1] FUSE is available at https://fusecorpus.github.io/FUSE/, with transcribed speech and facial expression data.

[2] The analysis in later sections also considered collected speech and galvanic skin response features.



*LREC-COLING 2024*, pages 7544–7555

2004). In contrast, this study used immediate post-task self- and partner-reported annotation, with speakers labeling their own and their dialogue partner's perceived emotional experiences, for rich insights. This approach taps into both perspectives (experiencing as well as perceiving) by asking participants to rate their own emotional experience, and also how they perceived their partner's emotional state.

In sum, we introduce the corpus **FUSE (FrUstration and Surprise Expressions)**, collected with speakers including dyads performing tasks, and analyze annotations in depth.

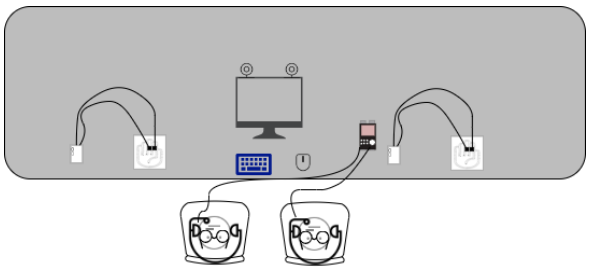


Figure 1: A speaker's voice was captured with a wearable microphone and TASCAM recorder, galvanic skin response (GSR) with a Shimmer3 on the non-dominant hand, and facial expressions using a webcam.

## 2. Relevant Prior Work

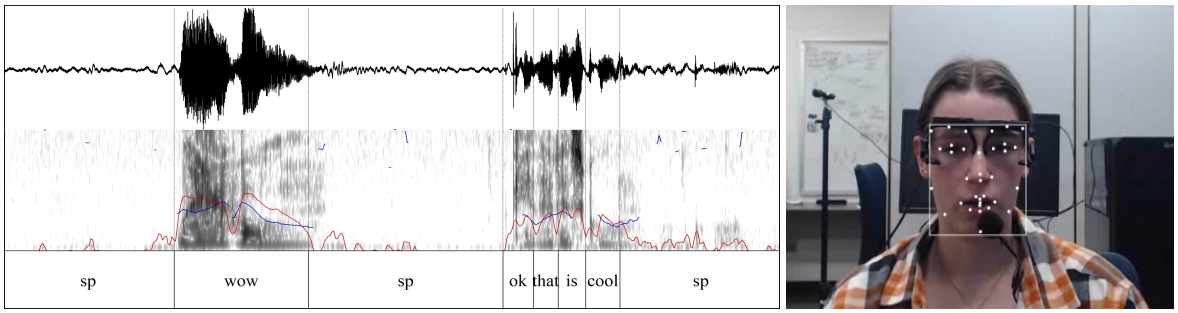
We focus on two emotions that enable human-AI collaboration or improve user experiences (Weidemann and Russwinkel, 2021; van der Burg and, 2022): *frustration* and *surprise*.

**Frustration** is related to anger (Berkowitz, 1990; Ang et al., 2002), but has its own characteristics. It occurs when there are barriers to achieving goals (Amsel, 1992). Prior works on frustration have used acted corpora (Busso et al., 2008) or lumped frustration together with anger or annoyance despite their distinctions (Schuller et al., 2010). From a use-inspired perspective, frustration is important to detect in human-computer interaction as it reveals troubling user experiences (Klein et al., 2002; Opoku-Boateng, 2015), and example applications include improving game and tutoring system designs (Gee, 2005; McQuiggan, 2007).

**Surprise** is a reaction that arises when a person faces sudden, unanticipated events (Meyer et al., 1997). Thus, experiencing the same event again will usually not re-surprise an individual. Surprise differs from other emotions in that it can have negative and positive polarity (Alm, 2010), and it is usually short-lived, but can impact the experiencer's subsequent behaviors and actions. A challenge in the elicitation of surprise data, is that it is also difficult to elicit many surprising reactions within a short time frame (Shea et al., 2018). Although surprise is often considered a basic emotion (Ekman, 1992), it has been scarcely studied in depth. The study of surprise can offer advantages in various domains, for example in human-AI teaming. From a human perspective, surprise can also aid comprehension (Loewenstein, 2019), cognition (Schomaker and Meeter, 2015), and memorability (Foster and Keane, 2018), thus promoting learning behaviors.

**Prior Emotion Corpora** In affective computing (Picard, 2000), researchers have studied the modeling of human emotion using datasets that leverage the Facial Action Coding System (FACS), which was based on early attempts to encode emotional expressions in facial muscle movements (Ekman and Friesen, 1978). There are multiple image/video-based corpora such as CK+ (Lucey et al., 2010) and AffectNet (Mollahosseini et al., 2017). Emotion recognition in text has been studied by the computational linguistics community and others for approximately two decades (Liu et al., 2003; Alm and Sproat, 2005; Alm, 2012; Mohammad, 2016), and examples of text-based emotional language resources include ANEW (Bradley and Lang, 1999) and GoEmotions (Demszky et al., 2020). Similarly, speech signal features have been used for speech emotion recognition (Abbaschian et al., 2021), research toward expressive speech synthesis (Alm, 2009), and other studies, using emotional speech corpora (Greasley et al., 2000; Engberg et al., 1997). However, rarely do corpora explicitly target frustration in particular, and approaches relying on unimodal representations are limited due to emotions' variation across speakers or dialogue interlocutors (Poria et al., 2017).

Multimodal corpora aim to provide comprehensive representations of human behavior (Shimojo and Shams, 2001), as in corpora such as RAVDESS (Livingstone and Russo, 2018) and MELD (Poria et al., 2019), which use combinations of speech, video, and text. ASCERTAIN (Subramanian et al., 2016) and DEAP (Koelstra et al., 2011) center on biophysical signals, and the HUMAINE Database (Douglas-Cowie et al., 2007) is a small resource with video clips. There is a scarcity in research on the multimodal analysis of frustration (Song et al., 2019) and surprise (Shea et al., 2018). We seek to fill this gap for studying surprise and frustration in multimodally grounded language. Finally, most corpora fail to cover dyadic interactions. While IEMOCAP (Busso et al., 2008) and SEMAINE (McKeown et al., 2011) do, our resource improves over IEMOCAP's acted interaction and SEMAINE's guided interaction by capturing spontaneous conversation.




(a) Speech spectrogram with transcription (b) Surprised facial expression

Figure 2: Panel (a) shows a speech signal spectrogram from an image task eliciting surprise, and the speech waveform in Praat (Boersma and Weenink, 2023). The red line indicates intensity and the blue line is the fundamental frequency associated with pitch. The third row shows the time-aligned utterances: *Wow. Ok, that is cool*, with *sp* representing a short pause. Panel (b) shows the participant's facial expression with eye widening and mouth open. The reported surprise rating by the subject for the stimuli was 4.87 of 5. In this stimulus, the participant was first asked about the sleep hours of ants and garden snails. Following a brief pause of 5 seconds, it was disclosed that ants have a sleep duration of merely 8 minutes within a 12-hour cycle, whereas snails are capable of sleeping for up to three hours. Subsequently, the subject provided a rating of 4.87 for the stimulus.

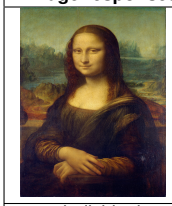
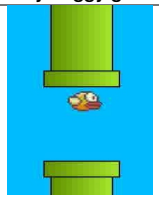
| **Image responses** | **Solve riddles** | **Play buggy game** |
|---|---|---|
|  | I speak without a mouth and hear without ears. I have no body, but I come alive with the wind. What am I? (Answer: An echo.) |  |
| individual | dyad | dyad |
| surprise | surprise | frustration |
| How **unexpected** was the image? Using the slider, indicate how unexpected the image was from 1 to 5, where 1 indicates totally expected. and 5 totally unexpected. | Using the slider, indicate the intensity for the **chosen emotion** from 1 to 5, where 1 indicates very low intensity and 5 very high intensity. | Choose the label that best describes how you believe your partner felt at the end of the game task? (Content, Frustrated, Surprised) |

Table 1: Three tasks to elicit surprise and frustration in speakers with representative rating questions.

## 3. Methods

**Data Collection Experiment** The data collection experiment used spontaneous elicitation. Three tasks, see Table 1, captured unscripted and natural emotional expressions in language and other modalities. Two tasks involved dyads engaging in dialogue and one was an individual task, for comparison. Each task session began with an introduction to the task. Then, participants engaged with stimuli eliciting emotional low-intensity reactions. The time window when the participant performed the task and experienced emotions was annotated by the participants with self-reported emotion labels. We recorded the speech, which was carefully transcribed, in addition to face video and skin response data. A diagram showing the data collection setup is in Figure 1. The tasks included:

1. **Replying to questions about images (an individual task)** Motivated by the use of visual clips or images in prior works (Uhrig et al., 2016), participants were shown nine image triads. First, a question prompted verbal responses. Then, an unexpected fact was revealed to elicit subtle surprise. For instance, participants were asked if they noticed anything unusual in an image of the Mona Lisa painting, and it was later revealed that the Mona Lisa did not have eyebrows. Next, participants rated their level of surprise on a Likert scale from 1 to 5, with 5 for highly surprising.

2. **Solving riddles (a dyadic task)** The riddle task asked participants to discuss four unexpected riddles, motivated by prior work that used riddles for surprise elicitation (Mahmoud et al., 2011). The task aimed to induce nuanced surprise in the interlocutors, who then annotated both their own emotions, choosing from *frustrated*, *surprised*, and *content*, intensity levels, and how they perceived the emotional reactions of their partners. The additional emotions, surprised and content, are important since frustration may combine with, for instance, amusement to form complex emotions.

3. **Playing a buggy game (a dyadic task)** Inspired by how a dysfunctional questionnaire induced frustration in a prior study (Hoque and Picard, 2011), we used a buggy version of the *Flappy Bird* game. Dyads took turns playing the game and coaching each other aloud, seeking a high score in a 3-minute session. The bugginess impeded scoring high, accumulating frustration while preventing boredom or fatigue

(D'Mello and Graesser, 2012). Afterward, participants self-annotated their emotions and how they perceived their partner's reaction, as in the prior task.

**Participants** The corpus has image task data from 19 participants, game task data from 39, and riddle task data from 38, with 19 participants completing all three tasks. The data included dyads (pairs) of speakers. Dyads completed the riddle and game tasks, and half also the image task. The task order was counterbalanced in pairs. The study took place in a university setting, and the participants included 19 female, 18 male, and 2 non-binary participants whose ages ranged from 18 to 35 years old, with the largest group being 19 years old. All subjects spoke English, with 28 identifying English as their L1 and 11 as their L2.

**Feature Extraction** For the analysis, we extracted features from four modalities: speech recordings, the transcribed spoken language, facial expressions, and GSR data. Prior studies have demonstrated the importance of speech features such as pitch and intensity for emotion recognition (Murray and Arnott, 1993). Additionally, linguistic features have been found to aid in affect modeling (Yoon et al., 2018). Time-aligned facial features (Zhi et al., 2020) and measured GSR (Ayata et al., 2016) serve as indicators of emotion too. Figure 2 shows an example of data from this study.

1. **Spoken Language Features** Transcripts for speech recordings were obtained using RevAI (Rev, 2022), Amazon Transcribe, and IBM Watson with word error rates (WER) of 15%, 22%, and 45%, respectively. These WER results reveal that ASR is brittle when faced with emotional speech and dialogue. Given its comparatively lower rate, we chose the RevAI transcription. In addition, to ensure quality corpus data, the transcripts were manually reviewed to correct any inaccuracies in the transcriptions. In addition, diarization of speaker turns was corrected (i.e., who was speaking when). We then used Sentence-BERT (Reimers and Gurevych, 2019) to generate sentence embeddings. We chose Sentence-BERT because it has been trained on both written and spoken text, considering its relevancy for encoding transcribed spoken language.

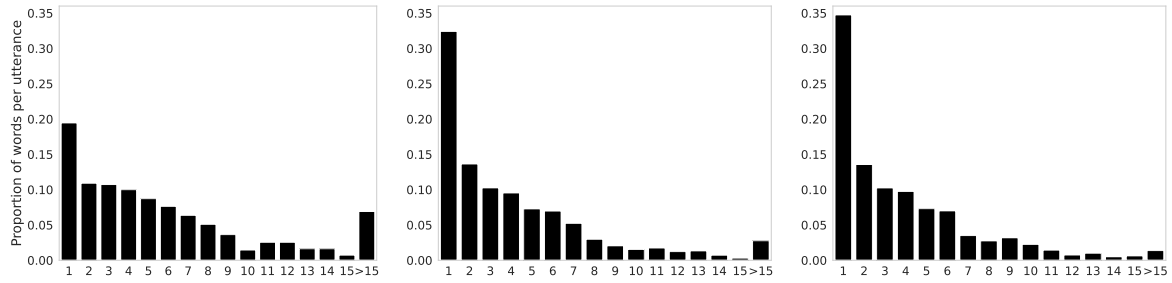
2. **Speech Features** Prosodic features such as F0 (a prominent contributor to what listeners perceive as pitch) and intensity are important for how speakers convey emotions, including surprise and frustration (Frick, 1986; Furnes et al., 2019), and also for listeners' perception of both emotions (Mozziconacci, 2002). Additionally, spectral features can also contribute to conveying these emotions (Ooi et al., 2014). Thus, we extracted Mel-frequency cepstral coefficients (MFCCs), pitch, and intensity from each speaker's speech signal (McFee et al., 2015), and applied Z-score normalization to account for speaker variation and gender differences.

3. **Facial Expression Features** Facial expressions were used to complement linguistic and speech signal features. Facial expressions can be important for conveying valence dimensions (Fagel, 2006). We used Affectiva's facial expression recognition model in iMotions (iMotions, 2015) to extract facial features such as attention, brow furrow, brow raise, cheek raise, chin raise, eye closure, eye widen, jaw drop, mouth open, smile, smirk, etc.

4. **Electrodermal Features** We used a Shimmer3 GSR+ (iMotions, 2021) to collect the electrodermal activity of biophysical skin response and consider the electrical conductance at various timestamps. One advantage of GSR is that it encodes a fast biophysical response, which can be useful for emotional reactions of subtle and short duration.

**Computational Experiments** Along with the label analysis, we also performed initial binary or tertiary emotion recognition modeling, considering both classical and well-defined neural classification methods. The class labels for the image task were *surprised* and *not surprised*. For the dialogue tasks, they were *surprised*, *frustrated*, and *content*. Per task, we defined a well-motivated *interval of emotion response*. The interval of emotion response corresponded to the time interval following the presentation of a stimulus during which the subject expressed their emotions and included verbal responses. For the image and riddle tasks, the interval of emotion response or the annotation region is fixed, while for the game-based tasks, the last 90 seconds were considered as the interval of emotion response because frustration accumulated over the course of gameplay.

Leveraging the extracted features, we explored both the early and late fusion of modalities (Boulahia et al., 2021), building a classifier for each task. For early fusion, we prepared ten data points from the predefined *interval of emotion response* for all four modalities. The choice of this quantity, specifically ten data points, was made to align with the average number of utterances per interval across all tasks. In the case of utterances, the initial ten utterances were selected. For all other modalities, if the number of data points within the interval exceeded ten, the data points were divided into ten segments and averaged to get ten data points, otherwise the last data point was duplicated to match ten. We used two methods: Random Forest (Buitinck et al., 2013) and an LSTM network




(a) Image task (n=550 utterances) (b) Game task (n=975 utterances) (c) Riddle task (n=1600 utterances)

Figure 3: The histograms show the proportion of word tokens per utterance in the three tasks, revealing that the riddle task elicited more utterances but a substantial proportion (approximately 35%) were one-word utterances. The individual image task elicited fewer utterances, while speakers tended to use longer utterances on average than in the dyadic tasks.

(Paszke et al., 2017) with two layers and a fully connected network to capture the temporal emotion data.

In the late fusion, we set the maximum sequence length to the $90^{th}$ percentile of the number of data points across each modality and built an individual network for each modality. Each network encoded each data point to a 200-dimensional vector. The outputs of four networks were stacked and passed through a fully connected neural network. Each individual network was a two-layered LSTM network followed by a fully connected neural network. We also experimented with transfer learning for speech features using EfficientNet B5 (Tan and Le, 2019). Since the size of the data was modest, we only considered the first 6 layers of EfficientNet features and froze their weights, and then connected the network to the adaptive average pooling layer and subsequently to a fully connected network. For meaningful comparison, we trained all the networks for 50 epochs as most networks converged within 50 epochs.

## 4. Results and Discussion

**Spoken Language Insights** An example utterance from a surprise task is in Figure 2, where two utterances, *Wow* and *Ok, that is cool*, are displayed. We also observe a rise in speech intensity and pitch on *wow*, suggesting a peak of surprise. This is consistent with a prior observation of surprised speech (Shea et al., 2018).

Table 2 shows utterance and word elicitation statistics by task. The individual task had a larger average utterance duration and average words per utterance compared to dyad tasks, likely due to adapting to turn-taking conventions between interlocutors in the dyadic tasks. The riddle task showed slightly shorter averages than the game. In addition, the histograms in Figure 3 also show that dyadic tasks had a higher proportion of one-word utterances compared to the individual task,

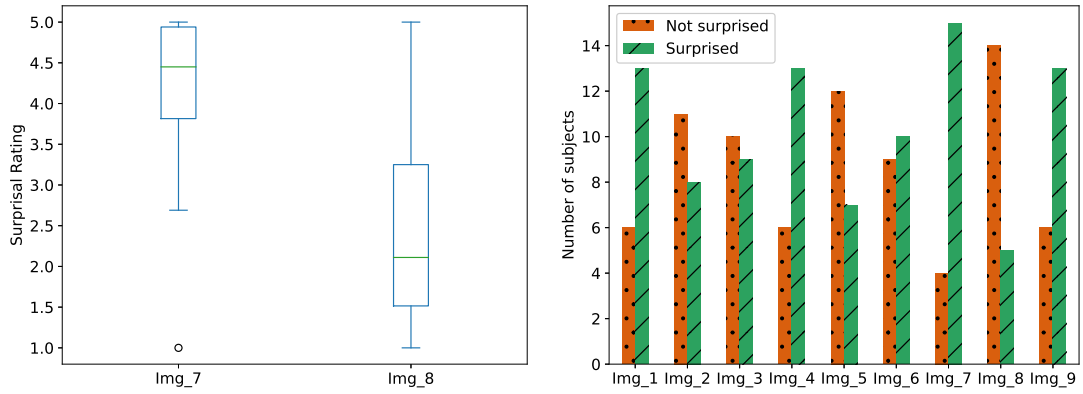
| | Avg utt. duration (s) | Avg words/utt. |
|---|---|---|
| **Image (ind.)** | 3.2 | 6 |
| **Riddle (dyad)** | 1.2 | 3.9 |
| **Game (dyad)** | 1.4 | 4.2 |
| **Dyad** | 1.3 | 4 |
| **All** | 1.6 | 4.4 |

Table 2: Utterance statistics per task. The mean utterance duration and mean words per utterance for the individual task are larger than for the dyadic tasks. Here, **Dyad** represents utterances from **Riddle** and **Game** combined while **All** represents the combined utterances from all tasks.

which likely reflects the absence of interruptions without a partner interlocutor, also supporting the shorter dyadic utterance durations in Table 2. In contrast, dyadic tasks elicited more utterances, which is expected because pragmatic meaning for task completion is co-created in dialogue by interlocutors. The findings also align with prior observations (Busso et al., 2008).

**Annotation Analysis** We also analyzed the annotations. The corpus consists of data from one individual and two paired tasks. Figure 6(a) and 6(b) indicate that the dyadic tasks elicited the intended emotions and also variation in emotion perception.

Figure 5 illustrates three comparisons analyzed for the dyadic tasks. The comparisons analyze the dyad interactions based on annotations. Let's refer to the two participants in each dyad task session as **A** and **B**. Following task completion, the participants annotate themselves, represented by **SA** and **SB**. Additionally, they provide annotations for each other, represented by **PB** and **PA**, based on their observation during the task. Comparison **I** analyzes the difference between participants' self-annotations and their annotations for their partners, that is SA versus PB and SB versus PA. Comparison **II** involves two comparisons: participants' self-annotation behaviors SA versus SB and their




(a) Contrasting image surprisal ratings.

(b) Label distribution across nine image stimuli.

Figure 4: Image task: Panel (a) shows box plots of two image stimuli having contrasting ratings, with one rated as eliciting high surprisal and another as low surprisal by the majority. Panel (b) shows the class distribution for nine image stimuli, showing that some elicited more surprisal responses.

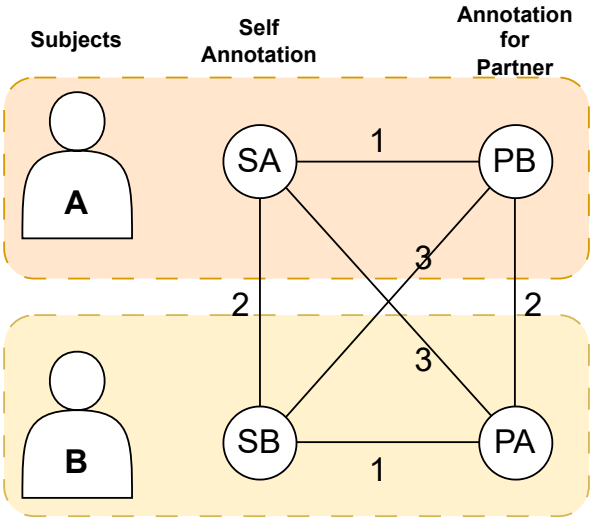


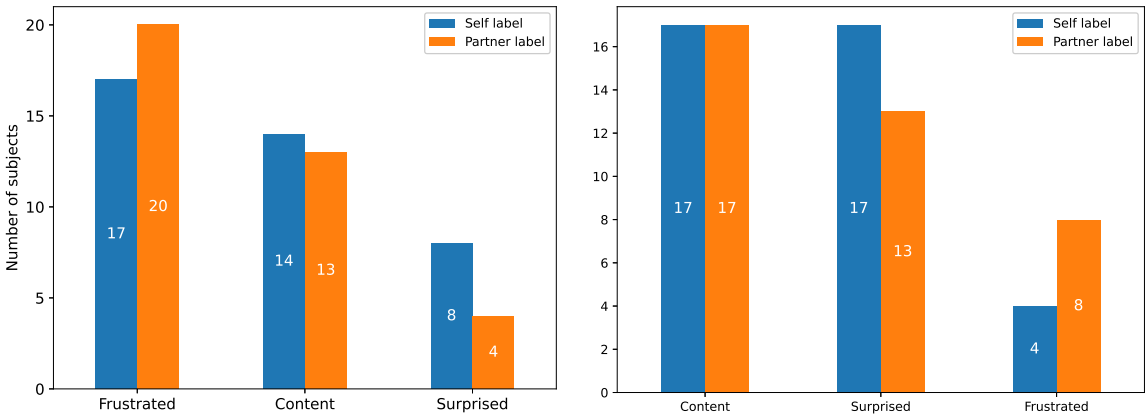

Figure 5: Annotation analyses considered self-and partner annotation. In a session with two speakers, speaker **A** reported their own emotion **SA** and how they perceived their partner's emotion **PB**. The other speaker **B** reported their own emotion **SB** and **A**'s emotion **PA**.

behaviors in annotating their partner's response PB versus PA. Comparison **III** explores how each participant annotated themselves compared to how their partners annotated them, that is, SA versus PA and SB versus PB.

1. **Individual Task's Annotations** In the Image task, each subject rated how surprised they were by each stimulus on a scale from 1 to 5, with 5 being the most surprised. The ratings were then converted to *not surprised or surprised* using a threshold (3.2) calculated from the mean of medians of two extreme stimuli in Figure 4(a). Panel 4(b) shows that the self-annotated surprisal varied by image.

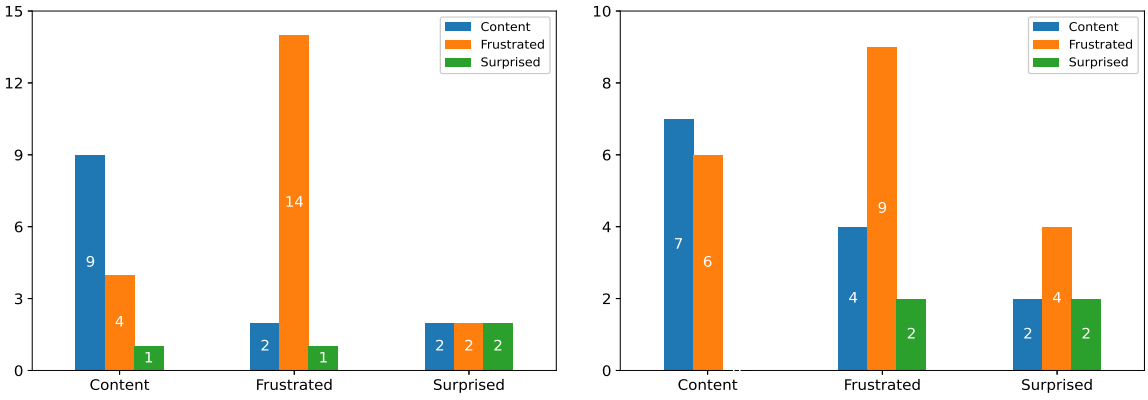
2. **Dyadic Tasks' Annotation** In the game and riddle tasks, participants selected both the emotion and intensity level that best described their emotional state and also reported how they perceived their partner's experience. The intensity reported by participants for themselves and their partner showed a Spearman correlation of $+1$, revealing that participants assumed their partner's emotional response had the same intensity level as their own–either for the same or another emotion. They mostly reported an intensity level of 3, suggesting that realistic low-intensity emotions were elicited, as intended.

3. **Game Annotations** Figure 7 shows two comparisons for the game task. Panel (a), which represents comparison I, suggests that when participants are *frustrated*, they perceive their partner as frustrated as well. However, if they are *surprised*, they are uncertain. Furthermore, when people are *content*, they often perceive their partner as content, but they may also perceive other emotions. A $\chi^2$ test examined the association between the reported emotions for comparison I, revealing a statistically significant association between self-reported and reported emotions for partners (p-value=0.007). For comparison II, $\chi^2$ tests revealed that there was no dependency (p-value=0.09) between how subjects annotate themselves in a session (SA versus SB) and even less dependency (p-value=0.25) between the labels annotated by others (PB versus PA). Panel (b) of Figure 7 shows comparison III or how speakers were perceived by their partners when they annotated themselves with certain emotion labels. A $\chi^2$ test showed that the association was not statistically significant. A visualization analysis suggested that when labeling themselves



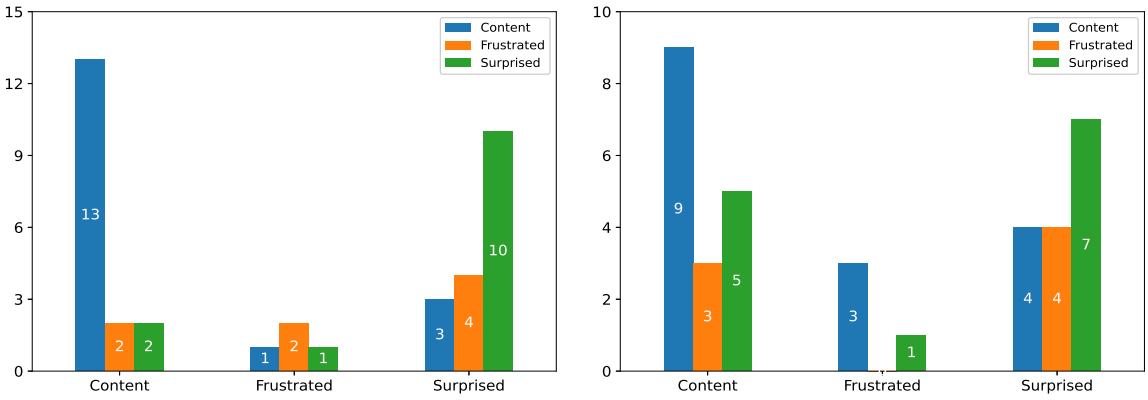

(a) Game: Self-labeling and partner-labeling. (b) Riddle: Self-labeling and partner-labeling.

Figure 6: (a) At the end of the **game** task, the majority reported themselves as *frustrated* and their partner too, indicating the game task elicited frustration. In addition, many felt *content* indicating blurring boundaries (e.g., frustration dealt with through humor). (b) For the **riddle** task, the majority reported feeling both *surprised* and *content*, and also attributed these emotions to partners. They labeled partners slightly more often as *content*.




(a) Comp. I: Partner-labels (bars), given self-labels. (b) Comp. III: Partner labels (bars), given self-labels.

Figure 7: **Game** task: Panel (a) shows that when speakers self-labeled as *frustrated* they mostly labeled their partner that too. Panel (b) shows that the speakers were often labeled as *frustrated* by their partner irrespective of what they self-labeled themselves, including when they self-labeled themselves as *frustrated* and *surprised*.




(a) Comp. I: Partner-labels (bars), given self-labels. (b) Comp. III: Labels by partner (bars), given self-labels.

Figure 8: **Riddle** task: Panel (a) shows that speakers mostly label their partner with the same label as themselves. Panel (b) shows that when speakers self-labeled as *surprised* or *content*, they mostly labeled their partner the same.

| Method | Early Fusion | | | | | | Late Fusion | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alg. | Random Forest | | | LSTM | | | LSTM | | | LSTM-TransferL | | |
| Task | Image | Game | Riddle | Image | Game | Riddle | Image | Game | Riddle | Image | Game | Riddle |
| **P** | 47 | 11 | 40 | 61 | 50 | 39 | 51 | 25 | 43 | 49 | 42 | 35 |
| **R** | 48 | 33 | 43 | 61 | 61 | 61 | 51 | 33 | 49 | 50 | 42 | 38 |
| **F1** | 47 | 17 | 41 | **61** | **49** | 39 | 51 | 28 | **46** | 49 | 42 | 36 |

Table 3: Performance for tasks across fusion methods: Based on the F1-score, LSTM with early fusion performed better than other methods for both the Image and Game tasks, while LSTM with late fusion showed improvement for the Riddle task. **P**, **R**, and **F1** represent Precision, Recall, and macro average F1-score metrics, respectively.

as *frustrated*, their partners largely tended to perceive them as such.

4. **Riddle Annotations** Panel 8(a), which illustrates comparison I, suggests that participants who felt *content* or *surprised* were mostly sure their partner felt the same. However, when participants reported feeling frustrated, they were not as certain about their partner's emotional state. As expected, frustration was reported infrequently for this task. Subjects mostly assign their partner the same label as themselves, as indicated by the significant $\chi^2$ test (p-value=0.005). This demonstrates inter-label dependencies for comparison I, which aligns with observations from the game task. Similar tests for comparisons II and III were not significant. Per panel 8(b), unlike the Game task, when participants reported feeling frustrated, their partner mostly perceived them as content but never as frustrated. A $\chi^2$ test for comparison III also indicated a statistically insignificant relationship; however, visual analysis suggests the presence of some patterns, warranting further investigation. In sum, people usually perceive others as feeling the same experience as themselves, and such annotation can be incongruent with how their partners annotated themselves.

**Computational Prediction Results** Table 3 shows computational classification results for the three tasks. The F1-scores indicate that early fusion improved predictions over late fusion for two out of three tasks (image and game) while late fusion did better for riddle data.

In addition, for early fusion, LSTM mostly performed better than Random Forest, except for the riddle task where results were comparable. LSTM provided substantial improvement for the game task. The boost by LSTM can be related to how LSTM can manage temporal data properties. Additionally, the late fusion used two networks: one with LSTM networks for all four modalities and another with LSTM network for three modalities and with an EfficientNet-based network for the speech audio modality utilizing the pre-trained weights of EfficientNet for transfer learning. The transfer learning method enhanced the F1-score for the game task substantially, by 14%.

The classification results indicate that the game is a challenging classification problem, and the image appears least challenging. We relate the game task results to the subtlety of frustration, and its potential to converge with other emotions into complex emotional states.

## 5. Conclusion

We introduced a novel, complex, yet interesting emotional language corpus – FrUstration and Surprise Expressions or FUSE. This corpus includes spontaneous, multimodally grounded language conveying subtle, realistic, and non-extreme frustration and surprise. It also provides individual and dyadic interactions. Analysis of annotation labels and spoken language yielded several insights, and results from predictive models were provided for the three tasks. We began to address the gap of low-intensity resources for emotions, as they tend to be frequent in natural settings, and this study explored frustration and surprise in collaboration contexts. Future work might compare modeling using self or partner ratings and explore contextual dependencies across subjects. Additionally, future analyses could conduct a qualitative examination of modeling results.

## 6. Ethics Statement

This work included human subjects research that was IRB-approved and involved informed consent. Participants were informed that they could withdraw from the study at any time. They received compensation in the amount of USD 15-25 for participating.

## 7. Acknowledgements

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this

material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## 8. Bibliographical References


Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249.

Cecilia O. Alm. 2009. *Affect in Text and Speech*. VDM Verlag.

Cecilia O. Alm. 2010. Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 118–122, Uppsala, Sweden. Association for Computational Linguistics.

Cecilia O. Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 107–112. Association for Computational Linguistics.

Cecilia O. Alm. 2012. The role of affect in the computational modeling of natural language. *Language and Linguistics Compass*, 6(7):416–430.

Cecilia O. Alm. 2022. *11 Linguistic data resources for computational emotion sensing and modeling*, pages 226–250. De Gruyter Mouton, Berlin, Boston.

Cecilia O. Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *Affective Computing and Intelligent Interaction*, pages 668–674, Berlin, Heidelberg. Springer Berlin Heidelberg.

Abram Amsel. 1992. *Frustration Theory: An Analysis of Dispositional Learning and Memory*. 11. Cambridge University Press.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of ICSLP*. Citeseer.

Değer Ayata, Yusuf Yaslan, and Mustafa Kamaşak. 2016. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In *2016 Medical Technologies National Congress (TIPTEKNO)*, pages 1–4. IEEE.

Lisa Feldman Barrett. 2004. Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, 87(2):266.

Leonard Berkowitz. 1990. On the formation and regulation of anger and aggression: A cognitive-neoassociationistic analysis. *American Psychologist*, 45(4):494.

Paul Boersma and David Weenink. 2023. Praat: Doing phonetics by computer [computer program].

Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech Report 1, Center for Research in Psychophysiology.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Sidney K D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):1–36.

Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. 2007. The HUMAINE

database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*, pages 488–500. Springer.

Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Inger S Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. 1997. Design, recording and verification of a Danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology*, pages 1695–1698.

Sascha Fagel. 2006. Emotional McGurk effect. In *Proceedings of The International Conference on Speech Prosody*, Dresden, Germany.

Meadhbh I Foster and Mark T Keane. 2018. The role of surprise in learning: Different surprising outcomes affect memorability differentially. *Topics In Cognitive Science*, 11(1):75–87.

Robert W Frick. 1986. The prosodic expression of anger: Differentiating threat and frustration. *Aggressive Behavior*, 12(2):121–128.

Desire Furnes, Hege Berg, Rachel M Mitchell, and Silke Paulmann. 2019. Exploring the effects of personality traits on the perception of emotions from prosody. *Frontiers in Psychology*, 10:184.

James Paul Gee. 2005. Learning by design: Good video games as learning machines. *E-Learning and Digital Media*, 2(1):5–16.

Peter Greasley, Carol Sherrard, and Mitch Waterman. 2000. Emotion in language and speech: Methodological issues in naturalistic approaches. *Language and Speech*, 43(4):355–375.

Mohammed Hoque and Rosalind W. Picard. 2011. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 354–359.

iMotions. 2015. Affectiva iMotions Biometric Research Platform. https://imotions.com/affectiva/. Accessed October 19, 2023.

iMotions. 2021. Shimmer3 GSR Sensor. https://imotions.com/products/hardware/shimmer3-gsr/. Accessed October 19, 2023.

Dacher Keltner, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, 43:133–160.

Jonathan Klein, Youngme Moon, and Rosalind W Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140.

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):18–31.

Jacqueline M Kory and Sidney K D'Mello. 2014. Affect elicitation for affective computing. *The Oxford Handbook of Affective Computing*, page 378.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pages 125–132.

Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Emotion recognition using multimodal deep learning. In *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23*, pages 521–529. Springer.

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):e0196391.

Jeffrey Loewenstein. 2019. Surprise, recipes for surprise, and social influence. *Topics in Cognitive Science*, 11 1:178–193.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101.

Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek. 2011. 3D corpus of

spontaneous complex mental states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 205–214. Springer.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Carol A McQuiggan. 2007. The role of faculty development in online teaching's potential to question teaching beliefs and assumptions. *Online Journal of Distance Learning Administration*, 10(3):1–13.

Wulf-Uwe Meyer, Rainer Reisenzein, and Achim Schützwohl. 1997. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21(3):251–274.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*, pages 201–237. Elsevier.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

John Morreall. 1982. A New Theory of Laughter. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 42(2):243–254.

Sylvie Mozziconacci. 2002. Prosody and emotions. In *Proc. Speech Prosody 2002*, pages 1–9.

Iain R Murray and John L Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108.

Michael Niepel, Udo Rudolph, Achim Schützwohl, and Wulf-Uwe Meyer. 1994. Temporal characteristics of the surprise reaction induced by schema-discrepant visual and auditory events. *Cognition & Emotion*, 8(5):433–452.

Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. 2014. A new approach of audio emotion recognition. *Expert Systems with Applications*, 41(13):5858–5869.

Gloria A Opoku-Boateng. 2015. User frustration in hit interfaces: Exploring past HCI research for a better understanding of clinicians'experiences. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1008.

Brian Parkinson, Agneta H Fischer, and Antony SR Manstead. 2005. *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology Press.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

Rosalind W Picard. 2000. *Affective Computing*. MIT press.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rev. 2022. RevAI: Speech to Text API. https://www.rev.ai/. Accessed March 19, 2024.

Judith Schomaker and Martijn Meeter. 2015. Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. *Neuroscience & Biobehavioral Reviews*, 55:268–279.

Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Jordan Edward Shea, Cecilia O. Alm, and Reynold Bailey. 2018. Contemporary multimodal data

collection methodology for reliable inference of authentic surprise. In *2018 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–5.

Shinsuke Shimojo and Ladan Shams. 2001. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509.

Meishu Song, Zijiang Yang, Alice Baird, Emilia Parada-Cabaleiro, Zixing Zhang, Ziping Zhao, and Björn Schuller. 2019. Audiovisual analysis for recognising frustration during gameplay: Introducing the multimodal game frustration database. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 517–523.

Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Meike K Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion elicitation: A comparison of pictures and films. *Frontiers in Psychology*, 7:180.

Vera van der Burg and. 2022. Ceci n'est pas une chaise: Emerging practices in designer-AI collaboration. In *Proceedings of DRS*. Design Research Society.

Alexandra Weidemann and Nele Russwinkel. 2021. The role of frustration in human–robot interaction – what is needed for a successful collaboration? *Frontiers in Psychology*, 12.

Neta Yitzhak, Nir Giladi, Tanya Gurevich, Daniel S. Messinger, Emily B. Prince, Katherine B. Martin, and Hillel Aviezer. 2017. Gently does it: Humans outperform a software classifier in recognizing subtle, nonstereotypical facial expressions. *Emotion*, 17:1187–1198.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.

Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. 2020. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36:1067–1093.