Human-Guided Modality Informativeness for Affective States

Torsten Wörtwein
twoertwe@cs.cmu.edu
Language Technologies Institute,
Carnegie Mellon University
Pittsburgh, PA, USA

Lisa B. Sheeber
lsheeber@ori.org
Oregon Research Institute
Eugene, OR, USA

Nicholas Allen
nallen3@uoregon.edu
Department of Psychology,
University of Oregon
Eugene, OR, USA

Jeffrey F. Cohn
jeffcohn@pitt.edu
Department of Psychology,
University of Pittsburgh
Pittsburgh, PA, USA

Louis-Philippe Morency
morency@cs.cmu.edu
Language Technologies Institute,
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

This paper studies the hypothesis that not all modalities are always needed to predict affective states. We explore this hypothesis in the context of recognizing three affective states that have shown a relation to a future onset of depression: positive, aggressive, and dysphoric. In particular, we investigate three important modalities for face-to-face conversations: vision, language, and acoustic modality. We first perform a human study to better understand which subset of modalities people find informative, when recognizing three affective states. As a second contribution, we explore how these human annotations can guide automatic affect recognition systems to be more interpretable while not degrading their predictive performance. Our studies show that humans can reliably annotate modality informativeness. Further, we observe that guided models significantly improve interpretability, i.e., they attend to modalities similarly to how humans rate the modality informativeness, while at the same time showing a slight increase in predictive performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**;
• **Applied computing** → *Psychology*.

KEYWORDS

multimodal, fusion, affective computing

ACM Reference Format:
Torsten Wörtwein, Lisa B. Sheeber, Nicholas Allen, Jeffrey F. Cohn, and Louis-Philippe Morency. 2021. Human-Guided Modality Informativeness for Affective States. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3462244.3481004

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10.
https://doi.org/10.1145/3462244.3481004

1 INTRODUCTION

Depression is a prevalent mood disorder that affects globally more than 264 million people [16]. Detecting depression early is crucial, as depression can affect the development of adolescents [4]. Therefore, we are interested in depression-related affective states during mother-adolescent interactions. To this end, we focus in this work on three affective states, i.e., positive, aggressive, and dysphoric, that have shown a relation to a future onset of depression [33] and study how these states are expressed through different modalities.

Combining information from multiple modalities to predict affective states is challenging and does not always improve predictive performance of machine learning models [40]. However, humans express themselves through multiple modalities, making it essential to study how humans integrate information from multiple modalities when recognizing affective states. Additionally, we are also interested in leveraging this knowledge to effectively combine information from multiple modalities in machine learning models. We focus our study on how humans use three important modalities for face-to-face conversations [22], i.e., vision, language, and acoustic modalities. While all three modalities may always be available, we hypothesize that a subset of modalities will be sufficient to predict expressions of affective states. In particular, we expect that these subsets are not the same for each instance of affect expression.

In this paper, we study the hypothesis of using a subset of modalities to predict affective states from two angels: (1) a human study to better understand which modalities people are paying attention to when recognizing affective states; and (2) the impact of integrating these human ratings to guide machine learning models to attend to a subset of modalities. An interesting aspect of this paper is that we are holistically studying the relation between modalities and affective states by showing annotators all available modalities at the same time and asking them to judge the informativeness of each modality. For these judgements, we discretize modality informativeness in three levels: (a) sufficient, when a modality is, by itself, enough to recognize the expressed affective state; (b) relevant, when a modality includes useful information about the expressed affective state but is not sufficient to recognize the affective state; and (c) none, when the modality does not seem to be used to express the affective state. We study whether human annotators can reliably accomplish this task and analyze the distribution of these modality informativeness annotations. Finally,

we explore the impact of integrating these annotations in predictive models. Our study and experiments are performed on a recent dataset of mother-adolescence interactions recorded in the context of studying affective states related to the onset of depression [24].

2 RELATED WORK

We group the related work in four topics. First, we cover computational approaches for predicting multimodal affective states. Then, we focus on how multimodal machine learning models estimate modality informativeness. Third, we mention multimodal perception experiments that highlight that affective states are differently perceived across modalities. Finally, we highlight some unimodal attempts at integrating human guidance to improve predictions of machine learning models, such as using eye gaze to attend to salient words in NLP tasks.

Multimodal affective recognition: A modality-centric view of affective states is to divide them into how they are expressed, i.e., non-verbally and verbally. Non-verbal affective states include, for example, the basic six emotions, while verbal affective states include more language-driven aspects of affect such as sentiment, complaining or (dis-)agreeing. Challenges such as AVEC [29], ComParE [32], and FERA [39] have focused extensively on predicting non-verbal expressions of affect. Similarly, language-driven aspects of affective states haven been predicted as part of sentiment analysis [34, 43] and to some degree as part of dialogue acts [5]. In this work, we focus on three multimodal affective states, i.e., they are expressed non-verbally and verbally, that have shown to be statistically related to a future onset of depression [33].

Modality informativeness: Many models that focus on multimodal fusion implicitly or explicitly estimate the informativeness of modalities [25, 37, 38]. Two motivations for modeling modality informativeness are often a better predictive performance [25, 37] and making the model more interpretable as the impact of each modality is estimated [38]. One way to model modality informativeness is to use modality attention with decision-fusion models [25]. As attention is not guaranteed to reflect how important a modality is [27], we guide the modality attention to be similar to human perceived modality informativeness and also evaluate how similar the predicted modality attention is to the perceived modality informativeness.

Modality perception: Multimodal perception studies have been conducted to rate how affective states are perceived in different modality combinations [3, 20, 23, 26]. As an example, some researches studied whether emotions are perceived differently across modalities [3]. Focusing on individual modalities has the advantages that other modalities cannot hinder the perception of the current modality, but being exposed to only a subset of modalities, i.e., not having all the available information, can lead to different judgments about affective states as demonstrated by these studies. To avoid this limitation and to focus instead on modality informativeness, we ask human annotators to judge modality informativeness while being exposed to all available modalities.

Human-guidance: Human attention, operationalized as eye gaze fixations, has helped in uni-modal tasks to learn more robust attention mechanisms in NLP as a way to attend to words [2]. In computer vision, eye gaze information was also used to attend to salient objects [36, 42]. While eye gaze is an effective way to derive visual attention, it is not well-suited to infer informativeness of other modalities such as for the acoustic modality. As an alternative, we ask human annotators to rate how informative each modality is.

3 DATASET

Our study takes advantages of the recent Transitions in Parenting of Teens (TPOT) [24] dataset which consists of 134 audio- and video-recorded mother-adolescent interactions (a total of 268 participants). These natural interactions are 20 minutes long and focus on problem-solving tasks. Conversations typically focus on discussing the amount of screen time, the participation in household chores, and the behavior towards other family members. All participating families are considered to have low social economic status in the US. The adolescents are between 11 and 14 years old, and half of the mothers have a history of a unipolar disorder.

Each interaction is annotated for four multimodal affective states: positive, aggressive, dysphoric, and other (mostly neutral). These affective states are closely related to the Living in Familial Environments (LIFE) codes [14] and can directly be derived from them [33]. The Krippendorff α of the annotated states is 0.66. The four affective states are expressed non-verbally and verbally. For example, being sad is coded as dysphoric but self-focused complaints are also coded as dysphoric.

The affective state coding focuses on the onsets of events, i.e., when enough evidence is available to determine an affective state. We assume that an annotated state is valid until the next onset. Through preliminary machine learning experiments, we determined that the annotations are most likely delayed by one second. We therefore shift all annotations by one second. The dataset has a total of 4,117 positive segments, 1,683 aggressive segments, 5,313 dysphoric segments, and 6,221 other segments. The average segment duration is 6.1 seconds.

4 HUMAN JUDGEMENT OF MODALITY INFORMATIVENESS

We are interested in how much information each modality contributes when recognizing affective states. Additionally, we want to explore whether interactions between modalities are crucial when predicting affective states or whether modalities can be used independently.

For our study, we recruited and trained two annotators from our local institution[1]. As the TPOT dataset contains sensitive data, all annotators were part of our IRB protocol. The annotation software ELAN [41] is used to display side-by-side videos of the mother and the adolescent. For each family member, we randomly select a balanced subset of twelve segments. We exclude segments of the "other" state from this annotation, as they are primarily characterized by neutral or no expressions. Appendix A provides more details on the annotation interface. Each video of a family member is randomly assigned to one of the two annotators. 10% of the videos (26

[1]The two female annotators were already familiar with the annotation software. We followed the established approach for annotator training where annotators are trained on a separate subset of the data (not used in our main study) until reaching a high enough agreement. In our case, we used the threshold of 0.7 Krippendorff α on the training subset.

videos) are annotated by both annotators to calculate the inter-rater agreement (see section 4.1 for details about the Krippendorff α).

Modality informativeness is defined as the amount of information a modality contains to recognizing an affective state. For each modality (vision, language, and acoustics), the annotators are asked "How much information does the modality contribute to the affective state?" and given the following response options: "sufficient information", "relevant information", "no information", and "not clear / I do not know". A modality is sufficient when the annotators can recognize the affective state using only this modality. In contrast, a modality is relevant, if by itself, this modality is not sufficient to recognize the affective state but it provides information towards the affective state. An example of relevant information is speaking loudly: it can signal a high arousal state, but typically we cannot differentiate between positive and aggressive states with just this cue. We should note that multiple modalities can be sufficient for the same segment. Furthermore, it is possible that none of the modalities is sufficient by itself, meaning that the interaction between modalities is crucial.

As a sanity check, we ask the annotators "Do you agree with the affective state?". This allows us to flag segments where the affective state might be too ambiguous. The annotators have the following response options: "agree", "somewhat agree (it could be interpreted as <affective state>)", "disagree", and again "not clear / I do not know". Our two annotators "agree" in 86% of the cases with the originally coded affective states. For our study, we exclude annotations where the annotators do not "agree" with the affective state.

4.1 Annotation Analysis

Annotator agreement: We report the agreement of our modality informativeness annotations using Krippendorff α: 0.50 for the visual modality, 0.66 for the language modality, and 0.65 for the acoustic modality. These Krippendorff α are computed using the ordinal weighting scheme [18] since our annotation label scheme is ordinal. If one or both annotators choose "not clear / I do not know" for a modality, we treat the annotation as missing. Only 6% of the modality annotations are flagged as missing, leaving 2,724 segments for vision, 2,694 for language, and 2,703 segments for acoustics (15.6% of all TPOT segments). While we sampled the affective states in a balanced manner, not all videos have three aggressive segments, leading to an imbalance between the affective states. Out of the segments that have at least one modality annotated (a rating different from "not clear / I do not know"), 35.55% are positive, 24.02% are aggressive, and 40.38% are dysphoric.

Modality Informativeness: We analyze the informativeness of each modality. As can be seen in Table 1, the vision modality provides most frequently sufficient information followed by the language modality. Interestingly, the acoustic modality does not seem to provide as much information for this dataset. A potential explanation might be that it is cognitively difficult to focus on acoustic characteristics when listening to speech [19]. It is further surprising to observe, that the annotators did not choose "relevant information" as often. This suggests that in most cases, individual modalities could be sufficient to predict an affective state.

Table 1: Distribution of the modality informativeness.

Modality	Information		
	No	Relevant	Sufficient
Vision	16%	11%	67%
Language	49%	3%	41%
Acoustic	78%	3%	13%

Table 2: Common behaviors related to the three affective states as reported by the annotators.

State	Behaviors
Positive	head node, yes / agree statements, smile, eyebrows raised, laughter
Aggressive	head shake, no / disagreement statements, scowl / glare, eyebrows raised, sigh
Dysphoric	gaze aversion, head facing downwards / away from partner, self-touches (face and head), fidgling, shoulder shrugs, lip suck/bite, sigh

While we did not annotate which exact behaviors are causing relevant/sufficient information, we asked our annotators for the most common behaviors for each of the three affective states and tabulate them in Table 2. Behaviors shared among affective states seem to be related to arousal (raised eyebrows) and valence (sigh). This is somewhat expected, since positive and aggressive states both tend to be high arousal states, while aggressive and dysphoric states both tend to be low valence states.

Informativeness and Missingness: The language and acoustic modality are not always available, since a person does not speak all the time. To validate if this has a big impact on the informativeness annotations, we look at how often words are spoken during segments that are annotated as containing "no information". If words are spoken during an uninformative ("no information") segment, we know that language and acoustics are available[2] and are not caused because speech is missing. During 51.15% of the uninformative language segments, words were spoken. Similarly for acoustics, 66.18% of the uninformative acoustic segments contain spoken words.

Modalities per affective state: Table 3 shows the distribution of informativeness for each affective state. Similar to Table 1, vision provides a lot of information across all affective states, but language provides more often information than vision for aggressive. In addition, language is more often informative for positive and aggressive than for dysphoric. A potential reason for this observation is that agreement and disagreement are coded as positive and aggressive, respectively. Another observation is that when the acoustic modality is informative, it tends to be informative for the positive state.

Cross-modal interactions: It is also interesting to study which modalities co-occur. Table 4 shows that more than half of the times

[2]This is a simplification for acoustics as people can also express themselves nonverbally, e.g., laughing, crying, and sighing.

Table 3: Percentage of available information for each affective state. 100% means all segments of the affective state.

Modality	Positive			Aggressive			Dysphoric		
	No	Rel	Suf	No	Rel	Suf	No	Rel	Suf
Vision	19%	6%	74%	24%	16%	55%	9%	12%	69%
Language	47%	2%	49%	25%	6%	63%	64%	3%	22%
Acoustic	70%	2%	26%	79%	9%	6%	83%	2%	4%

Table 4: Co-occurrence of available information (relevant or sufficient). Co-occurrence probabilities are relative to how often the row modality is informative, e.g., in 38% of the cases when vision is informative, language is also informative.

Modality	Co-occurs with		
(base rate)	Vision	Language	Acoustic
Vision (83%)	100%	38%	18%
Language (48%)	67%	100%	9%
Acoustic (17%)	89%	26%	100%

when language is informative, vision also provides information. When the acoustic modality is informative, it is often accompanied by visual information. While a single modality is frequently sufficient, affective states are often still expressed in multiple modalities. A predictive model could benefit from this extra information in terms of robustness by integrating uni-modal predictions dynamically based on a predicted modality informativeness.

5 MODALITY ATTENTION

To guide how much attention a model pays to each modality, we decide to explore two decision-fusion architectures that differ only in how modalities are aggregated. The first architecture averages unnormalized predictions (logits) while the second architecture averages normalized predictions (probabilities). While unnormalized logits contains more information than the normalized probabilities, the weighting of the unimodal predictions (attention) might be misleading as the unimodal models can learn to encode modality informativeness through the magnitude of their unnormalized logits instead of relying on the attention mechanism [27].

We use superscript in the following equations to denote a modality $m \in M$ with $M = \{v, l, a\}$. The prediction $\hat{y}_i$ of the unnormalized model for segment i is expressed as

$$p_i = \text{softmax}\left(\left[\sum_{m \in M} l^m_{i,\text{Pos}} a^m_i, \ldots, \sum_{m \in M} l^m_{i,\text{Oth}} a^m_i\right]\right) \quad (1)$$

$$\hat{y}_i = \arg\max_{s \in \{\text{Pos, Agg, Neg, Oth}\}} p_{i,s} \quad (2)$$

where $[\cdot]$ is the concatenation operator. The unnormalized logits $l^m_i \in \mathbb{R}^4$ for each modality m are defined as

$$l^m_i = W^m f^m(X^m_i) + b^m \quad (3)$$

and the attention vector $a_i \in \mathbb{R}^{|M|}$ is

$$a_i = \text{softmax}\left(g([\, f^v(X^v_i), f^l(X^l_i), f^a(X^a_i)\,])\right) . \quad (4)$$

W is the projection matrix to the four affective states and b is the bias term. f and g are operationalized using Multi-Layer Perceptrons (MLP). This first model is part of the family of cooperative gating models [15] and is a special case of the multimodal gating unit [25] when used on the predicted output.

The second model averages normalized probabilities. The changes to the first model are

$$l^m_i = \text{softmax}(W^m f^m(X^m_i) + b^m) \quad (5)$$

$$p_i = \left[\sum_{m \in M} l^m_{i,\text{Pos}} a^m_i, \ldots, \sum_{m \in M} l^m_{i,\text{Oth}} a^m_i\right] . \quad (6)$$

5.1 Human-Guided Attention

Our goal is to study how models can be guided to prioritize modalities similarly to how humans judge the modality informativeness. Maximizing this similarity has the potential advantage of better interpretability and could also help the model during training to focus on the subset of informative modalities as it might prevent the model from learning some spurious correlations.

To improve the similarity between model attention and human judgments, we propose a new auxiliar loss. To formalize this loss, we define two matrices $A, H \in \mathbb{R}^{n \times |M|}$ where n is the number of segments. These matrices correspond to the predicted attentions (A) and the human informativeness judgements (H). Row i in these matrices corresponds to the importance of the three modalities for segment i. To define a similarity between the human judgements and the algorithmic attentions, we convert the ordinal human judgments to numeric values: no information (0.0), relevant information (0.5), and sufficient information (1.0).

$A_m, H_m \in \mathbb{R}^n$ are the columns of A and H respectively. They correspond to attention values of modality m across all n segments. We minimize the following auxiliar loss

$$-\lambda \frac{1}{|M|} \sum_{m \in M} \text{pearson}\,(A_m, H_m) . \quad (7)$$

This loss maximizes the modality-averaged correlation between A and H. $\lambda \in \{0.1, 0.5, 1.0\}$ is a hyper-parameter to find a good scale for the auxiliar loss.

6 EXPERIMENTAL METHODOLOGY

To evaluate our human-guided prediction approach[3], i.e., the unnormalized model with the auxiliar loss from Equation 7 (referred to as guided), we define two baseline models: the normalized and unnormalized model each without the auxiliar loss (referred to as normalized and unnormalized respectively).

We define an interaction-independent five-fold split for testing with a nested holdout split for validation (60% for training, 20% for validation, and 20% for testing). Reported metrics are averaged over the five test sets. The following hyper-parameters are validated for all models: learning rate for Adam [17], number of layers of the individual MLPs, strength of the L2-norm for the learnable parameters, and λ to balance the auxiliar loss. The primary loss

[3]The code is available at github.com/twoertwein/HumanGuidedAttention.

function is the categorical cross-entropy, rectified linear units are used as non-linear activation functions, and early stopping is used. All parameters are jointly optimized. Features are z-normalized using the respective training sets and feature selection is performed with a linear support vector classifier [11] on the training sets. The best model is determined by the weighted accuracy averaged over the validation sets.

We report the affective state prediction performance using accuracy (Acc) and Krippendorff α[4]. Krippendorff α is chosen since we can easily compare the model's performance with the inter-rater agreement.

Further, we use two metrics to evaluate how interpretable the learned attention is. For each modality m, we report

$$\rho_m = \text{spearman}(A_m, H_m)\ . \tag{8}$$

Compared to section 5.1, we replace the differentiable Pearson's correlation with the non-differentiable Spearman's correlation since the human informativeness scale is ordinal. Additionally, the human informativeness and the predicted attention should for each segment have a similar/the same ordering. We report the segment-averaged Spearman's rank correlation coefficient

$$\bar{\rho}_{\text{local}} = \frac{1}{n}\sum_{i=1}^{n} spearman(A_i, H_i) \tag{9}$$

to evaluate whether segments have on average a similar attention ordering as the human informativeness. $A_i, H_i \in \mathbb{R}^3$ are the rows of A and H.

Significance tests are conducted with paired person-clustered bootstrapping [28] using $p = 0.05$ and 10,000 resamplings at the person-level.

6.1 Extracted Features

Vision: We use OpenFace [1] and AFAR [7] to extract facial action unit intensities and occurrences, head rotation, and eye gaze angles. When aggregating frame-level features to the labeled segments, we ignore features of video frames that according to OpenFace/AFAR were not correctly tracked. All features are aggregated to the labeled segments using the mean and standard deviation. Additionally, the maximum is used when aggregating facial action unit intensities. As an additional proxy for gaze aversion, we calculate the angular distance from looking straight into the camera [13] as the camera is located approximately on face-level behind the conversation partner.

Language: All interactions are manually transcribed. Words are automatically aligned to the audio using the Montreal Forced Aligner [21]. We use the dimensions from LIWC 2015 [35] to represent all words that occur during a labeled segment.

Acoustic: The audio files are first processed with StereoTool's declipper[5] in an attempt to recover clipped amplitudes caused by a too high microphone gain and then volume-normalized with FFmpeg according to the EBU R128 standard. Next to features from COVAREP [6], we extract the feature sets corresponding to the following openSMILE [10] configurations: eGeMAPS v01a [8], prosodyAcf (pitch and voicing probability), and vad_opensource [9] (speech activation detection). Most acoustic features are meaningful only while a person is speaking. When aggregating the audio features to the labeled segments, we consider audio features that happen only while speaking according to the aligned transcripts and when COVAREP/openSMILE detect speech. All low-level features are aggregated to the labeled segments using the mean and standard deviation. The high-level features from eGeMAPS are aggregated using only their mean.

[4]Krippendorff α is typically computed between the ratings of annotators. Here, we treat the model and the ground truth as two raters.
[5]www.stereotool.com

Table 5: Performance on the entire test set and the gating metrics on the annotated test subset.

Model	α	Acc	$\bar{\rho}_{\text{local}}$	ρ_v	ρ_l	ρ_a
Chance	0.000	0.307				
Normalized	0.336	0.528	0.284	0.273	0.356	-0.148
Unnormalized	0.350	0.537	0.372	0.140	0.288	-0.090
Guided	**0.351**	**0.541**	**0.636**	**0.288**	**0.423**	**0.283**

Table 6: Performance on the annotated test subset. Oracle refers to using the annotated modality informativeness instead of the learned attention.

Model	Krippendorff α	Accuracy
Unnormalized	0.318	0.518
- Oracle	0.324	0.535
Guided	0.328	0.525
- Oracle	**0.379**	**0.561**

Table 7: Average of the predicted attention for the three annotated affective states on the entire test sets.

Modality	Positive	Aggressive	Dysphoric
Vision	0.607	0.542	0.689
Language	0.329	0.422	0.276
Acoustic	0.064	0.036	0.035

7 RESULTS AND DISCUSSION

Human-guided attention: Results are summarized in Table 5. Our human-guided model shows small improvements over the baseline models but most importantly the learned attention weights are much closer to human judgement. The correlation between the attention and the human judgement significantly increased from $\bar{\rho}_{\text{local}} = 0.372$ to $\bar{\rho}_{\text{local}} = 0.636$ meaning that our guided model prioritizes modalities similar to how humans prioritize them. The modality-specific correlations (ρ_v, ρ_l, and ρ_a) increased as well, making it easier to interpret the attention across segments.

Oracle experiment: Table 6 shows the hypothetical case when our guided model predicts perfectly the human informativeness. Its performance would significantly improve from $\alpha = 0.328$ to $\alpha = 0.379$. The other models do not improve significantly when using the human informativeness.

Attention per modality and affective state: Finally, Table 7 shows the averaged attention of our guided model for the three

annotated affective states on the test sets. It is very intriguing to compare Table 7 and Table 3. This comparison shows similar trends between the human judgement and the model's attention: vision is essential and language is more important for positive and aggressive than for dysphoric. The only obvious difference is that the model amplifies the existing bias [44] of acoustics not being too predictive.

8 CONCLUSION AND FUTURE WORK

This paper studied the hypothesis that a subset of modalities is sufficient to recognize affective states from two perspectives. First, we demonstrated that humans can reliably judge the informativeness of modalities and observe that in most cases a single modality is sufficient to recognize affective states while at the same time the affective states are still expressed through multiple modalities. Second, we proposed a human-guided auxiliary loss to improve the learned attention to be significantly more similar to human informativeness judgements while not degrading the predictive performance. Finally, the predictions can further be improved by directly using the human informativeness judgments during test time demonstrating empirically that the human ratings are reliable. This paves the way for more intuitive and easier to interpret multimodal models.

Achieving a significant improvement when overwriting the learned attention with the human judgement indicates that our model can be corrected by a trained human which makes our model more controllable and potentially also more acceptable by users [30]. This significant improvement also highlights the need for more research on how to learn better and more robust attention mechanisms.

Affective states encompass a variety of phenomena [31]. Affective states that are defined differently than the three studied affective states will most likely have a different modality informativeness. Further, some behaviors are more prominent in interactions between unfamiliar people than between familiar people, for example smiles [12], which might also shift the modality informativeness depending on the conversational setting.

In our annotation study, annotators were simultaneously exposed to three modalities. This makes it challenging to entirely ignore the influence of other modalities when judging how informative a single modality is. We encourage future work that contrasts modality informativeness when judging modalities independently and judging them jointly.

ACKNOWLEDGMENTS

This material is based upon work partially supported by the National Science Foundation (Awards #1722822 and #1750439), and National Institutes of Health (Awards #R01MH125740, #R01MH096951, #U01MH116925, and #U01MH116923). Any opinions, findings, 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 or National Institutes of Health, and no official endorsement should be inferred.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 302–312.

[3] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[4] Alison L Calear and Helen Christensen. 2010. Systematic review of school-based prevention and early intervention programs for depression. *Journal of adolescence* 33, 3 (2010), 429–438.

[5] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*. 225–234.

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[7] Itir Onal Ertugrul, Laszlo A Jeni, Wanqiao Ding, and Jeffrey F Cohn. 2019. Afar: A deep learning based tool for automated facial affect recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–1.

[8] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[9] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller. 2013. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 483–487.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[11] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research* 9 (2008), 1871–1874.

[12] Allison S Gabriel, Jennifer D Acosta, and Alicia A Grandey. 2015. The value of a smile: does emotional performance matter more in familiar or unfamiliar exchanges? *Journal of Business and Psychology* 30, 1 (2015), 37–50.

[13] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing* 32, 10 (2014), 641–647.

[14] Hyman Hops, Betsy Davis, and Nancy Longoria. 1995. Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system. *Journal of Clinical Child Psychology* 24, 2 (1995), 193–203.

[15] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[16] Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 392, 10159 (2018), 1789–1858.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

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

[19] Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology* 54, 5 (1957), 358.

[20] Sydney Lolli, Ari D Lewenstein, Julian Basurto, Sean Winnik, and Psyche Loui. 2015. Sound frequency affects speech emotion perception: Results from congenital amusia. *Frontiers in Psychology* 6 (2015), 1340.

[21] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.. In *Interspeech*, Vol. 2017. 498–502.

[22] Albert Mehrabian. 1971. *Silent messages*. Wadsworth Publishing Company, Belmont, California.

[23] Gelareh Mohammadi, Sunghyun Park, Kenji Sagae, Alessandro Vinciarelli, and Louis-Philippe Morency. 2013. Who is persuasive? The role of perceived personality and communication modality in social multimedia. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 19–26.

[24] Benjamin W Nelson, Lisa Sheeber, Jennifer Pfeifer, and Nicholas B Allen. 2021. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry* 62,

2 (2021), 199–211.
[25] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated Multimodal Units for Information Fusion.. In *ICLR (Workshop)*.
[26] Emily Mower Provost, Yuan Shangguan, and Carlos Busso. 2015. UMEME: University of Michigan emotional McGurk effect data set. *IEEE Transactions on Affective Computing* 6, 4 (2015), 395–409.
[27] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4782–4793.
[28] Shiquan Ren, Hong Lai, Wenjing Tong, Mostafa Aminzadeh, Xuezhang Hou, and Shenghan Lai. 2010. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics* 37, 9 (2010), 1487–1498.
[29] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 3–12.
[30] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.
[31] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.
[32] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. 2011. The INTERSPEECH 2011 speaker state challenge. (2011).
[33] Orli S Schwartz, Michelle L Byrne, Julian G Simmons, Sarah Whittle, Paul Dudgeon, Marie BH Yap, Lisa B Sheeber, and Nicholas B Allen. 2014. Parenting during early adolescence and adolescent-onset major depression: A 6-year prospective longitudinal study. *Clinical Psychological Science* 2, 3 (2014), 272–286.
[34] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
[35] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[36] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. 2012. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 91–98.
[37] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
[38] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2020. NIH Public Access, 1823.
[39] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. 2017. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 839–847.
[40] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multimodal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12695–12705.
[41] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.
[42] Fen Xiao, Liangchan Peng, Lei Fu, and Xieping Gao. 2018. Salient object detection based on eye tracking data. *Signal Processing* 144 (2018), 392–397.
[43] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.
[44] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989.

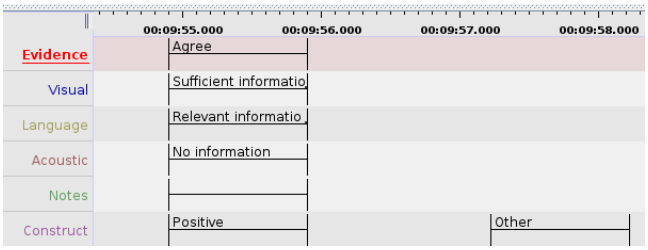

Figure 1: The annotation interface. On the left an annotated onset and on the right a nearby onset for context. The length of the segments has no meaning. The onset is the start of the segment.

A ANNOTATION INTERFACE

Figure 1 shows a screenshot of the annotation interface. The side-by-side videos are located above the tiers (not shown in the screenshot). Knowing that another onset happens immediately before or after an onset that is going to be annotated, was pointed out to be important by our two annotators during pilot studies. To contextualize the sampled onsets, nearby onsets are included in the ELAN files if they occur within five seconds.