Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild

Vargas-Quiros, Jose; Cabrera-Quiros, Laura; Oertel, Catharine; Hung, Hayley

DOI
10.1109/TAFFC.2023.3269003

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Vargas-Quiros, J., Cabrera-Quiros, L., Oertel, C., & Hung, H. (2023). Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild. *IEEE Transactions on Affective Computing*, *15*(2), 519-534. https://doi.org/10.1109/TAFFC.2023.3269003

Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild

Jose David Vargas Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung

***Abstract*—Although laughter is known to be a multimodal signal, it is primarily annotated from audio. It is unclear how laughter labels may differ when annotated from modalities like video, which capture body movements and are relevant in in-the-wild studies. In this work we ask whether annotations of laughter are congruent across modalities, and compare the effect that labeling modality has on machine learning model performance. We compare annotations and models for laughter detection, intensity estimation, and segmentation, using a challenging in-the-wild conversational dataset with a variety of camera angles, noise conditions and voices. Our study with 48 annotators revealed evidence for incongruity in the perception of laughter and its intensity between modalities, mainly due to lower recall in the video condition. Our machine learning experiments compared the performance of modern unimodal and multi-modal models for different combinations of input modalities, training, and testing label modalities. In addition to the same input modalities rated by annotators (audio and video), we trained models with body acceleration inputs, robust to cross-contamination, occlusion and perspective differences. Our results show that performance of models with body movement inputs does not suffer when trained with video-acquired labels, despite their lower inter-rater agreement.**

***Index Terms*—Action recognition, annotation, continuous annotation, laughter, laughter detection, laughter intensity, mingling datasets.**

I. Introduction

LAUGHTER is traditionally associated to its characteristic vocalization (ie. the sound of laughter). In research too, its vocal manifestation has received the most emphasis.

Nonetheless, laughter is a multimodal phenomenon. Darwin presented a curious depiction of excessive laughter: ”the whole body is often thrown backwards and shakes, or is almost convulsed; the respiration is much disturbed; the head and face become gorged with blood; with the veins distended; and the orbicular muscles are spasmodically contracted in order to protect the eyes. Tears are freely shed.” [1, p.208]. This depiction makes reference to multiple characteristic manifestations of laughter: the facial movements of laughter, the full-body movements of laughter, and the physiological changes of laughter.

Following this premise, works in social signal processing [2], [3] have delved into the problem of automatically detecting and classifying laughter from audio, video and audiovisual recordings of its manifestations. Annotation is a key step in these studies. The first step in annotation of naturally occurring laughter usually involves the temporal localization, or segmentation of laughter (from its context). Next, laughter segments or episodes are categorized or otherwise rated. Functional or formal categorizations are the most common, but no consensus coding schemes exists for either of these tasks. Laughter intensity is also a common variable of interest that has been rated in multiple studies [4], [5], [6], [7], [8], [9], [10]. Mazzocconi et al. [11] have linked laughter intensity directly to the meaning of laughter, as an indication of the magnitude of a positive shift in arousal caused by the laughable (the object of laughter) in the laughing subject.

Nevertheless, the emphasis on the vocal manifestations of laughter translates strongly to its annotation, where laughter has most commonly been annotated from audio or audiovisual face recordings, by third-party observers [12], [13], [14], [15]. Less commonly, laughter has been annotated from body movements alone, using video. This has been done in in-the-wild datasets of *mingling crowds* recorded in real-life events [16], such as the dataset in Fig. 1. In these datasets, audio recordings are commonly not available, due to the technical and logistic difficulty, and privacy challenges when equipping each study participant with a microphone. In-lab studies of the body movements of laughter have also often opted for video-only annotation of laughter, to align with the target task under study.

However, it is unknown if video labelling of laughter has a relevant effect on annotation quality, and how annotations acquired in this way differ from the more common audio-based and audiovisual annotations. The same is true for audio-based labeling: the benefits of including video at annotation time have not been verified. In other words, the consequences of the choice of annotation modality have received little attention in research. Furthermore, it is unclear whether ratings of intensity of laughter

Manuscript received 15 November 2022; revised 25 February 2023; accepted 6 April 2023. Date of publication 25 May 2023; date of current version 23 May 2024. This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant 639.022.606. Recommended for acceptance by F. Ringeval. *(Corresponding author: Jose David Vargas Quiros.)*

Jose David Vargas Quiros, Catharine Oertel, and Hayley Hung are with the Department of Intelligent Systems, TU Delft, 2628 Delft, The Netherlands (e-mail: j.d.vargasquiros@tudelft.nl; c.r.m.m.oertel@tudelft.nl; h.hung@tudelft.nl).

Laura Cabrera-Quiros is with the Escuela de Ingenieria Electronica, Instituto Tecnologico de Costa Rica, Cartago 30109, Costa Rica (e-mail: l.c.cabreraquiros@tudelft.nl).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2023.3269003, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2023.3269003

Fig. 1. Screenshots from the four elevated views in our dataset of free-standing interactions used in this work.

can be expected to be congruent across modalities, a question with direct implications in the interpretation of laughter [11].

While inter-rater agreement is an important dimension of annotation quality, higher annotation agreement does not necessarily imply superior model performance. The question of how annotation modality impacts model performance is therefore a separate, yet also unexplored question.

Answering these questions is important both for the interpretation of previous work focusing on a single modality, and for informing annotation choices in future work. In this work, we take a first step in that direction by studying laughter annotation across modality conditions. First, we investigate how human ability to detect, segment and estimate intensity of laughter (three foundational tasks in laughter work) differ with and without access to video or audio. Due to the difficulty of collecting audio in in-the-wild mingling settings in particular [16], we use an in-the-wild mingling dataset containing full-body motion information. Data was collected during a real-life event, and contains naturally-occurring laughs (Fig. 1). Body movements of laughter (eg. shaking, swaying, arm and feet movements) can be observed in the videos, but access to facial features is limited due to occlusion. These factors, along with the diversity of camera angles, and distances to the camera make the in-the-wild mingling setting one of the most challenging scenarios for laughter perception, especially from video. It should however be noted that our dataset showcases a specific range of conditions and our answers may not generalise to other scenarios with, for example, more consistent access to facial or body movements.

Second, we study how labels acquired under different modality conditions affect the performance of machine learning models for laughter detection, segmentation and intensity estimation. We pay special attention to the question of whether video-acquired annotations result in performance comparable to that of audio and audiovisual annotations. Naturally, the input modality of the model itself plays an important role here. We compare models with the same input modalities used in annotation: video, audio and audiovisual. Additionally we included accelerometer readings from chest-worn wearable devices (worn by many subjects in our dataset) as an additional model input. Such accelerometer readings have been used in previous work for the detection of multiple social actions such as speaking [17], [18], [19], with performance competitive and often superior to that of video. Furthermore, wearable accelerometers have privacy and scalability advantages due to their low cost and their ability to capture information from the device wearer alone. We hypothesized that acceleration would have a behavior similar to video, since both modalities capture primarily body movement information. However, we expected acceleration to better capture laughter intensity when compared to video due to its orientation invariance, and to it not being affected by occlusion and cross-contamination like video is. Our contributions are the following:

- We present the first human study of laughter annotations across annotation modalities, comparing between three conditions of interest in previous work: audio-only, video-only and audiovisual. We studied the three annotation tasks of laughter detection, time-localization and intensity rating.
- We present a cross-modal analysis of annotations via inter-annotator agreement within and between annotation conditions: video-only, audio-only and audiovisual. We obtained insights important both for the interpretation of previous work annotating on a single modality, and for informing annotation choices in future work.
- We investigated the effect of annotation modality in machine learning model performance. Mirroring the human study, we used state-of-the-art models for detection, intensity estimation and time-localization. We implemented, trained and evaluated models for different combinations of input modalities (audio, video, acceleration), training and testing label modality (video, audio and audiovisual). It is shown that despite the lower inter-annotator agreement of video-based labeling, they may be entirely appropriate to train models for laughter detection from body movements.

II. Background and Related Work

In this section we discuss laughter annotation in research, especially in computational work towards understanding laughter. In Section II-A) we start by briefly summarizing part of the research landscape surrounding laughter. In Section II-B we discuss automatic laughter detection and related machine learning tasks. In Section II-C we discuss work on laughter annotation and how laughter has been annotated in previous studies.

A. The Study of Laughter in Interaction

Laughter has been approached from the perspective of multiple scientific disciplines. Psychology and linguistics are concerned with, among others, the semantics and functions of laughter in interaction [20], [21], [22], [23], [24], [25]. In biology, the evolutionary role [26], [27] and physiological effects of laughter [28] are subject of study. Meanwhile, social signal processing, speech and human-agent interaction fields are concerned with automatic tasks such as laughter detection [29], [30], classification [31], [32] and synthesis [33], with datasets being created for the study of laughter in specific [13], [34], [35].

Laughter is most often analyzed as a meaningful signal in social interactions, as it is an overwhelmingly social phenomenon found to be about 30 times more likely in social situations than when by oneself [36]. To this end, drawing a parallel with the study of speech, Mazzocconi et al. [11] distinguished two broad levels for the study of laughter: 1) laughter form and context and 2) laughter's (social) meaning and function. Laughter form includes the physiology and body movements of laughter, and its acoustic features; and laughter context includes the situation in which it is produced [37], its positioning with respect to speech, to others' laughter, and to its object (the laughable) [36], [38]. Most of the work on the form of laughter is concerned with it's phonetics and acoustic structure, with different coding schemes for segmentation of laughter into its constituent (acoustic) components often being used [39]. Laughter intensity has also received attention as a dimension of laughter form [4], [5], [6], [7], [8], [9], [10], [40], [41]. Most laughter in conversations has been observed to occur at relatively low intensity [11], [41].

Laughter form and context influence its second level of analysis: the meaning and function of laughter. Mazzocconi et al. proposes the following as the meaning of laughter: "The laughable l having property P triggers a positive shift of arousal of value d within A's emotional state" [11, p.4], where A is the producer of the laugh. This interpretation provides a link between laughter intensity and laughter meaning. Despite the importance of laughter intensity in previous work, it is not known to what extent intensity ratings are congruent (or not) across modalities.

Laughter has been found to serve a multitude of functions at the coordination level as a cue for topic termination [42], [43]) and at the social level to foster relationships, cooperation and group cohesion [27], [44].

B. Automatic Laughter Detection, Classification, and Intensity Estimation in the Wild

Most research in laughter detection and classification has made used of meeting datasets and focused on the audio and audiovisual modalities. Truong et al. [14], [45] used spectral features, pitch, energy and voicing to discriminate laughter from speech. In a series of papers, Petridis et al. investigated audiovisual laughter detection and discrimination [46], [47] from upper body meeting videos, using static and dynamic features fed into a single-layer perceptron. Bohy et al. [48] studied laughter and smile classification from audiovisual recordings, including the role of intensity levels.

There have been fewer attempts to automatically assess laughter exclusively from the video modality. Mancini et al. [4] proposed a method to estimate laughter intensity from the movement of shoulder and head keypoints in a video. More recent action recognition methods based on 3D convolutional neural networks (CNNs) [49] have not been applied and analyzed in this task.

Full body poses and acceleration have also been inputs of interest. [32], [50] showed that traditional classifiers are capable of recognizing and classifying elicited laughter from pose sequences alone.

The related task of voice activity detection (VAD) has received more attention in in-the-wild settings, with models having been proposed for the detection of speech from video alone [51], [52]. Here, a deep 3D-CNN-based model has been shown to improve over previous approaches [53]. Additionally, work with accelerometer inputs has shown that this modality holds sufficient discriminative power to improve over larger video-based methods [18], [53].

C. Laughter Perception and Annotation

At its lowest level, laughter annotation is concerned with the recognition and segmentation of the form of laughter. Most of the work on the form of laughter is concerned with it's phonetics and acoustic structure. Laughter is typically classified in voiced, unvoiced and speech laughter (speech with laughter characteristics) [54]. Distinction between voiced and unvoiced is often made based on the degree of engagement of the vocal chords [55]. Speech laughter, although traditionally receiving less attention than isolated laughter, has been found in some cases to have comparable frequency of occurence [56]. Regarding its temporal extent, there is not a widely-accepted definition of what constitutes a laughter episode. Most studies of laughter delving into its structure have relied on audio waveforms for the segmentation of laughter, typically into laughter syllables or vowels (ha) at the lowest level, followed by bouts (sequences of syllables), which are separated by inhalations [57]. Truong et al. propose a multi-level segmentation scheme to describe the structure of laughter, including speech laughter [39]. This scheme, however, relies on audio alone.

Body movements, especially those occurring below the face, have been largely disregarded in the study of laughter form. There are however, notable exceptions. In a perception study, Griffin [32] showed that humans are capable of recognizing laughter and even of classifying it functionally based on stick figures. The use of stick figures provided a way to isolate the body movement component of laughter. Note however, that in contrast to our work, this study was not concerned with annotation (where the goal is to use the most reliable information available) and did not analyze agreement across modalities. Bohy et al. [58] studied the correlations between 2D joint displacements and audio laughter intensity and Hammoudeh [59] found differences in the body movements of laughter between males and females.

In the work most similar to ours, but focused exclusively on facial movements, authors created visual, audio, and audiovisual laughter stimuli/examples from face recordings [60]. The audio contained different levels of artificial noise to make laughter more difficult to detect. 20 annotators indicated if they perceived laughter or not in these examples. The goal was to study how much the face contributes to the perception of laughter. The study reported that "visual laughter consistently made auditory laughter more audible" (ie. audiovisual laughter was easier to detect than audio-only laughter), a phenomenon also observed previously for speech perception [60]. Although this is, to the best of our knowledge, the only work to perform a cross-modal analysis of laughter perception, its findings do not necessarily generalize to our setting, where the video modality contains

overall body movement information, but facial movements are not consistently available. Furthermore, being a perception study, they considered expert annotations to be ground truth, but provided no analysis of inter-rater agreement.

Most studies of automatic laughter detection (see Section II-B) rely on laughter annotations made from audio [30], [46] (possibly automatic like the ones in AMI [61] and SEMAINE [62]) or audiovisually [50], [63], [64]. However, studies concerned with the body movements of laughter often obtained ground truth annotations from the video modality alone. [4] rated laughter intensity from body movements alone. Cu et al. [65] annotated five affective categories of laughter from body movement, without sound. These studies, however, do not offer a comparison with audio-based annotation, and it is therefore uncertain to what extent annotations would be congruent across modalities.

III. Our Approach

Answering our research questions requires the annotation of a large set of laughter segments with associated audio and video signals. Measuring inter-annotator agreement across conditions additionally requires that the same segments are annotated by multiple annotators. Annotations must also be done by a representative sample of annotators, large enough to ensure that individual biases do not drive the results. The first step in a study of laughter in the wild is to localize laughter in the target dataset. Ideally, a large number of annotators would each watch our complete dataset (with more than 50 h of individual behavior) to find and annotate laughter episodes. This, however, would involve thousands of hours of human labour. Due to the relative scarcity of laughter in conversation in the wild, most of this time would be spent listening to speech with only sporadic laughter. Therefore, the first simplification that we adopted was to pre-localize *laughter candidates*. *Laughter candidates* are segments (or thin slices) where the author of the study (who did the pre-annotation) perceived laughter to occur, with some temporal context around the laugh (details in Section V-A). The pre-annotation was done inclusively, meaning that in case of doubt laughter was always annotated. These positive candidates were complemented with negative examples, where laughter was not perceived to occur, to obtain a dataset of laughter/non-laughter candidates. The resulting dataset was used both for human annotation and machine learning experiments.

Fig. 2 shows an overview of our study. In Section IV we present the audiovisual dataset chosen. In Section V we delve into our methodology for: the design of our human annotation study (Section V-B); analysis of annotator agreement (Section V-C); and analysis of machine learning model performance (Section V-D) for classification, segmentation and laughter intensity estimation.

IV. Dataset

Our dataset was collected during a business networking event in Delft, The Netherlands. Subjects in the experiment were members of a group organizing regular events. During the event, most of the interaction consisted of free-standing conversation (Fig. 1). Participants were free to move around as they pleased. While the event also included several pre-planned activities including a social game and music performance, we excluded these moments and made use only of the segment of the data containing free interaction. The following data was collected during the interaction:

Body Acceleration: A wearable accelerometer sensor that was hung around the neck like a badge measured upper torso acceleration. Importantly, the weight of our device was such that it in most situations it would not swing erratically but rather rest against the chest.

Individual Audio: Lavalier-type microphones attached to the face of participants via Lavalier tape recorded sound at 44.1 kHz. Microphones were connected to a Sennheiser SK2000 transmitter worn around the waist area. Our audio equipment consisted of 32 such microphones. These individual audio recordings were used to obtain Voice Activity Detection (VAD) labels at 100 Hz for each participant, making use of rVAD [66], a state-of-the-art unsupervised VAD method specially designed for noisy audio. 100 Hz is the fixed output frequency of rVAD and enough to capture even single syllables in languages like English [67].

Video: 12 overhead cameras and four side-elevated cameras were placed above and in the corners of a video zone. Participants were informed about this video zone, and asked to stay outside if they did not wish to be recorded. In this work we only make use of the side elevated cameras, due to it being a viewpoint more familiar to observers and able to capture the face. Fig. 1 shows a capture of the four elevated camera views.

In coordination with event organizers, it was decided that each participant would be free to choose which sensors to wear (microphone, accelerometer, or both). Of about 100 attendees to the event, 43 wore a sensor during the event. Of them, 20 were male and 23 female. The rest decided not to take part of the data collection, or could not be given a sensor due to our supply limit.

While similar *mingling* datasets have been published in the past [16], [68], our dataset was the first to contain high-quality individual audio recordings, opening the door for cross-modality studies such as this one.

V. Methods

In this section, we detail the methods used in our study of laughter for: 1) obtaining laughter/non-laughter candidates for annotation, 2) laughter annotation, 3) the study design (ie. assignment of laughter candidates to annotators, and related decisions) and 4) automatic laughter assessment.

A. Laughter Candidate Generation

To obtain the previously introduced laughter candidate segments to be annotated in our human study, the authors localized any *possible occurrences* of laughter in the dataset by watching the audiovisual recordings for every data subject and segmenting perceived laughter episodes using the annotation software

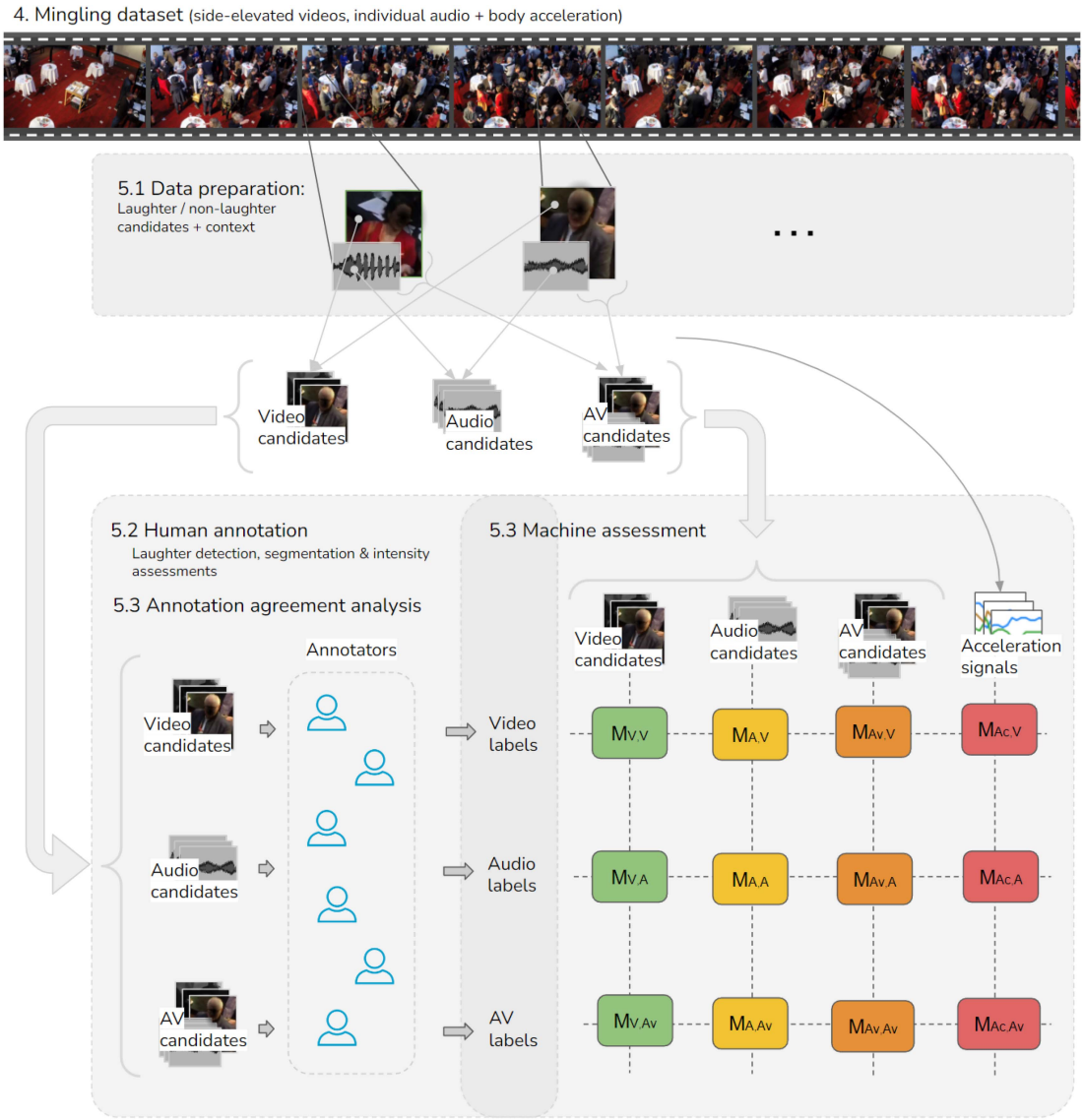
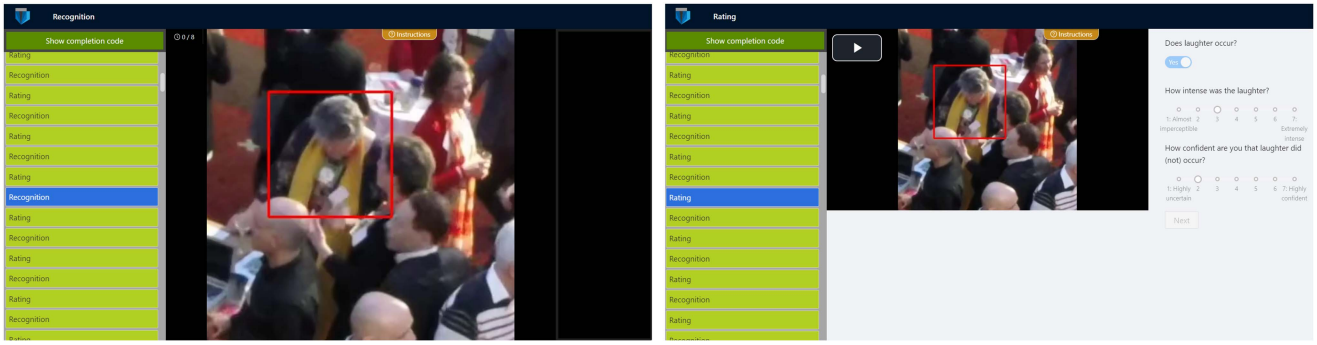


Fig. 2. Overview of our study. From a mingling dataset with video, individual audio, and accelerometer readings (Section IV), we extracted pre-annotated segments of potential laughter and speech, each of $7\,s$ in length. These segments were annotated for laughter presence, segmentation and intensity under three conditions: audio-only, video-only and audiovisual. We analyze the labels directly (Section VI-A) and use the different sets of labels to train and benchmark models for laughter detection, segmentation and intensity estimation (Section VI-B).

ELAN [69] to indicate the start and end of each laugh while referencing the audio waveform (normal ELAN annotation process). We were deliberately inclusive by annotating segments when in doubt and we localized all types of laughter, including speech laughter. We included cases where subjects appeared to be laughing in the video but the audio was not clear, or vice-versa. The cameras in which a particular laughter episode was visible were also annotated. Segments not visible in at least one of the videos were discarded. Segments present in multiple cameras were only considered once by randomly picking one of the cameras. Finally, annotations closer than $1\,s$ apart were joined into a single laughter episode. Regarding duration of the laughs, onset and offset inhalations [11] were not considered to be part of the laugh, since they were most often hard to perceive among the cocktail party noise in the scene. At this stage, our candidate set consisted of 459 laughter candidates. Next we complemented this set with segments likely to be negative (no laughter).

1) Negative Candidate Generation: As negative samples we extracted a number of segments likely containing no laughter from the rest of the dataset. To avoid having mostly segments of *listening behavior* in this negative set (our conversing groups were often large), we sampled negative candidates from speech utterances as given by our VAD labels. Additionally, since some data subjects were much more likely to laugh than others, we sampled the distribution of negative samples per subject proportional to the distribution of positive samples. Concretely, our sampling procedure is:

1) samples a subject S with probability $P_L(S)$ where $P_L(S)$ is the probability of a positive laughter candidate belonging to subject S.



(a) Recognition and continuous annotation step. (b) Intensity and confidence annotation.

Fig. 3. Screenshots of the annotation interface in Covfee [70]. The two steps shown were repeated for every example that an annotator rated. In (a) annotators were shown a target person marked by a red box, and part of the scene around the target, and instructed to hold down a key when laughter was perceived to be occurring by the target person. The interface provided visual feedback when the key was held down. In (b), subsequently, annotators rated laughter intensity (Likert scale 1–7) and their confidence in their assessment (Likert scale 1–7).

2) samples a speech utterance uniformly from the set of speech utterances of S of length $l_{\min} < l < l_{\max}$ where $l_{\min}$ and $l_{\max}$ are the lengths of the shortest and longest laughter episodes. This was done to avoid very long speech utterances from being introduced as negative examples.

We aimed for 30% of total candidates (about 200) to be negative, to provide a large-enough set for computing laughter detection agreement, but maintaining a majority of positive examples in which laughter intensity could also be rated. This resulted in a total of 659 candidates (459 positive and 200 negative).

2) Expanding Candidates in Time: Finally, laughter candidates (positive and negative) were expanded in time. The goal was to more closely resemble the process of annotating laughter in the wild, where it is unknown when laughter might happen, and allow the annotator to understand some of the context of the scene. To this end, we expanded each segment at both ends with a duration randomly (uniformly) sampled between 1.5 s and 3.5 s. We set the bottom of this range (1.5 s) to be close to the mean length of a laughter episode. Empirically, this was enough to process the scene and be ready for annotation. We set the top of the range (3.5 s) with the goal of obtaining total segment lengths below 10 s to maintain the annotation process fast. We used a uniform distribution to minimize predictability of the location of the laughter episode.

3) Spatial Localization Via Bounding Box: Since our side-elevated camera views captured most of the interaction scene, the target subject needed to be extracted or indicated to annotators. This was done by annotating a single, tight, bounding box around the target person for the first frame of the video. To allow annotators to use visual context of the scene while providing good visibility of the target subject, videos were cropped beyond the borders of this bounding box by multiplying its width and height by 3 (constrained to fitting within the frame) and maintaining its center. Our observations showed that this was in most cases enough to capture the interlocutors of the target person. The target person's box was shown to the annotators before the start of the video (see Fig. 3(a)).

B. Annotation of Laughter Candidates

Central to our study of laughter annotation is the design of the process to be followed by annotators. The first step in annotation of laughter in the wild is the localization of the laughter episodes in time. This process may range from spontaneous annotation of laughter with little instruction to annotators, to carefully-directed segmentation following a specific protocol. We leaned towards spontaneous, less instructed annotation due to it being a common first step in the annotation of in-the-wild datasets. Note that careful segmentation of laughter boundaries may not be necessary for many uses of such datasets where achieving a high time precision is not the goal. Furthermore, precise segmentation may be performed as a second step over roughly-localized laughter instances. Our goal was therefore not to obtain precise separation of individual laughter events such as episodes or bouts, but rather to obtain usable indications of laughter occurrence.

Actions are traditionally localized in videos using tools such as ELAN [69], where the user localizes the start and end frame of the action by drawing an interval on top of an audio waveform. In tools such as Vatic and CVAT, actions are annotated via flags, which are turned on for the frame when the action is deemed to start, and off for the end of the action. In affective computing, *continuous annotation* techniques are commonly used to annotate variables such as arousal and valence. In *continuous annotation*, annotators control the value of the target variable while watching the subject in video, usually without pause. This has the advantage of letting the annotator perceive the behavior without interruption, and being efficient and predictable in terms of time needed to annotate. On the other hand, continuous methods also necessarily introduce a reaction time delay. Multiple techniques have been proposed to mitigate these delays.

We chose to make use of continuous annotation for our study due to the mentioned advantages. We mitigated annotation delay by making use of an experimentally defined offset, as detailed in Section V-B3. We made use of a binary action localization technique implemented in the Covfee framework [70], which

asks annotators to hold down a keyboard key when they perceive laughter to be occurring. Its graphical interface is shown in Fig. 3(a). This process allows annotators to maintain focus on the videos by minimizing the input effort, while still giving us access to high-resolution segmentations of laughter. Since the annotation time is shortened and predictable, this process also allowed us to obtain more annotations per annotator, relevant for our study design (Section V-B1).

After the continuous annotation step, for each candidate, we asked annotators explicitly whether they perceived laughter to occur, their perceived laughter intensity, and their confidence in their laughter ratings (Fig. 3(b)). Annotators could replay the laughter episode if they desired.

1) Crowd-Sourced Annotation Process: As introduced in Section V, answering our research questions requires annotations of laughter under three conditions: audio-only, video-only and audiovisual. Measuring agreement within a condition imposes the requirement that at least two annotators rate each $(candidate, condition)$ combination. A sufficient number of candidates must also be annotated to be able to train our computational models and measure agreement over a large-enough set. Finally, for access to a large pool of annotators, annotations would be crowd-sourced and each annotation HIT (human intelligence task) should ideally not last longer than approximately 45 minutes to avoid fatigue. In our tests, we estimated each candidate to require about 30 seconds for annotation. This imposes an upper bound on the number of samples per annotator of around 90, which we reduced to 84 to have room for error.

One other natural choice to consider was whether to use a between-subjects or within subjects design. To maximize the number of annotators per condition, we opted for a study where each annotator takes part in all three conditions. To avoid bias, we impose the restriction that one annotator never annotates the same candidate under different (or the same) modalities, ie. one annotator rates three disjoint sets of candidates.

According to these design decisions, we divided our 659 candidates into 7 sets of 84 examples and one set containing the remaining 71 candidates. Each of these candidates sets was in turn divided into three equal-size subsets (for the three conditions). Each permutation of these three subsets resulted in a different human intelligence task (HIT), each containing the same candidate subsets but mapped to different conditions. Fig. 4 is a diagram of this process for each set of 84 candidates. Each HIT was completed by two annotators. Annotating all candidates required 48 annotators in total. This design allowed us to compute pair-wise inter-annotator agreement (per condition) over sets of 28 paired ratings, for 24 distinct pairs of annotators. This resulted in a total of 3954 annotator ratings.

2) Annotation HITs: We crowd-sourced our annotations to 48 annotators via the Prolific crowd-sourcing platform [71]. We implemented the complete annotation flow using the Covfee annotation framework [70]. Each HIT contained several introductory tasks and examples, followed by three annotation blocks, one for each modality condition. The order of video-only, audio-only and audiovisual blocks was randomized to avoid ordering bias due to factors like fatigue. The ordering of laughter examples within each block was also randomized for the same

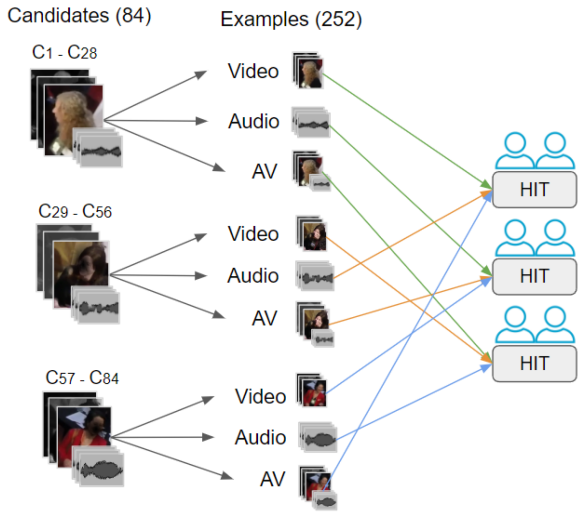


Fig. 4. Structure of the annotation stage of our study. Sets of 84 randomly-selected candidates are separated into 3 equal-size sets of 28 candidates. Candidates are then separated into their audio and audiovisual modalities and assigned to HITs such that each HIT contains 28 distinct candidates per condition. Each HIT was annotated by 2 annotators.

reason. The detailed structure of a HIT is presented in Appendix A.1, which can be found on the Computer Society Digital Library.[1] Statistics of the ratings provided by each annotator, time to complete the experiment and experience ratings are presented in Appendix A.2, available in the online supplemental material.

3) Annotation Delay Correction: Delays in continuous annotation have been investigated within the affective computing community for continuous-value annotations of affective dimensions. Some works have proposed machine learning models that are robust to annotation delays [72], [73]. Marioryad et al. [74] proposed a method for correcting delay by maximizing mutual information between the continuous label time series and an auxiliary signal containing facial keypoints. However, the authors also showed that simply offsetting annotations by a constant value resulted in performance comparable to that of more complex schemes.

Despite these results, it is unclear to what extent delay depends on the particular actions being annotated. We therefore decided to measure delay directly for our task and annotators. At six points in each annotation HIT (two per condition, see Section V-B2), we inserted special *calibration* (positive) laughter examples, which were the same for all annotators. We precisely labeled the onset and offset times of laughter in these six examples, using ELAN [69]. This allowed us to calculate a delay in the annotator's continuous labels, to approximate the average delay of each annotator. We used this average annotator delay as correction offset for an annotator's labels.

C. Measuring Inter-Annotator Agreement

We designed our study for the computation of inter-rater agreement, or reliability, within and across conditions. Cohen's

[1]Online available: https://doi.org/10.1109/TAFFC.2023.3269003

Kappa, Fleiss' Kappa, and Krippendorfs Alpha are some commonly used measures of agreement. For nominal values (eg. laughter/non-laughter) Cohen's Kappa is capable of computing agreement between exactly two annotators. Although Cohen's Kappa is subject to biases in some instances, it still has been recommended by previous work for fully crossed designs with multiple coders, by computing the average of pairwise agreement [75]. Since each of our annotator groups rated a set of examples not rated by any other pair (ie. our study consists of a set of fully-crossed designs), we used this approach to measure agreement for nominal values.

Cohen's Kappa is however not appropriate for interval/ordinal values like laughter intensity (Likert scale 1–7). Here, we used Krippendorff's alpha, a reliability measure applicable to any number of raters and which adjusts for small sample sizes. We averaged pairwise Krippendorff's alpha values over rater pairs.

Measuring agreement between time series is a more complex challenge. Straightforward application of agreement measures like Cohen's Kappa at the frame level fails to consider the strong dependencies between contiguous frames. Measures specifically designed for segmentation such as Staccato [76] and Gamma [77] are not suitable to our use case due to their assumption that every annotated segment corresponds to a distinct event, separate from any contiguous units. Our goal was not to obtain precise separation of individual laughter events such as episodes or bouts, but rather to obtain good indications of when laughter was occurring. Given this goal, we consider a measure such as Intersection over Union (IoU) to be appropriate because it indicates the degree of overall overlap between annotations. We computed IoU as the size of the intersection of positive annotations over the size of the union of positive annotations. This is inspired on action localisation metrics used in computer vision [78], which make use of IoU to identify true positives. An IoU of one indicates perfect correspondence between the positive sections of the time series, while an IoU of zero indicates no overlap. In cases where both signals contain only negatives (IoU is undefined), we set the metric to one to indicate full agreement. It should be noted that IoU does not correct for chance agreement and is therefore not an inter-rater agreement measure such as Krippendorff's alpha, which can be compared across datasets. However, since our goal is to observe potential differences across modalities over the same set of underlying data, we do not require such correction.

D. Automatic, Laughter Detection, Intensity Estimation, and Segmentation

Video-based models for detecting, assessing (eg. intensity) and segmenting actions have been extensively studied in computer vision and pattern recognition (Section II-B). We make use of modern approaches within these fields. Regarding the video modality, due to the small size of our dataset, training state-of-the-art methods from scratch would be infeasible. We focused on approaches with pre-trained models available to use as feature extractors. Among those, 3D convolutional neural networks (CNNs) are known to reliably achieve top performances in action recognition benchmarks. We decided to make use of a 3D ResNet pretrained on Kinetics-400, a large action recognition dataset with 400 action classes and over 300000 labeled video clips. The network implementation and models are available as part of the *Pytorchvideo* library [79].

Regarding audio-based models, work by Gillick et al. [29] investigated laughter detection in two datasets with significant background noise. One of these, the Audioset dataset [80] is freely available to download. This dataset of 10-second clips from Youtube videos recorded in a variety of in-the-wild settings contains 5696 clips labeled as containing laughter. In their implementation, the authors provided a list of randomly-sampled no-laughter clips to complete the dataset with negative. Given that this dataset had more examples and variety of subjects than ours, we decided to pre-train the audio-based model on it. We made use of the same model proposed by Gillick et al. [29]: a 2D ResNet model operating on the spectrogram of the audio inputs. We trained on 85% of the dataset, with 15% separated to determine a good stopping point. We otherwise used the same hyper-parameters used by the authors.

As motivated in Section III, we made use of acceleration as an additional modality capturing body movements. As acceleration-based model, we made use of a ResNet variant for time series, implemented as part of the *tsai* library [81]. Given the much lower dimensionality of the acceleration data (compared to video and audio), and the lack of availability of comparable acceleration datasets, we trained this model from scratch.

For both video and audio models, we used pre-trained models as feature extractors by freezing all parameters and removing network heads. For classification, the features output by the base networks (with dimensionalities: 2304 (audio), 8192 (video) and 128 (acceleration)) are fed into a head consisting in a linear layer followed by an output sigmoid layer and binary cross-entropy loss, standard choices for binary classification. For multimodal evaluation, we concatenate the features from multiple models before the head of the network.

We decided to approach intensity estimation as a regression task, given the interval nature of the ratings. We follow the same model structure, but we removed the sigmoid computation from the output and used L2, or mean squared error (MSE) loss, a standard choice for data with no outliers.

For segmentation, we decided to approach the task as the estimation of a binary mask (ie. of our continuous binary annotations). This would allow us to use the same base networks and pre-trained models. However, multimodal fusion should now be done earlier, since the time dimension encodes information likely useful for segmentation. We therefore implemented separate segmentation heads per modality, which are fused at the output via average pooling. For all models, we apply pooling and convolution operations over the spatial and channel dimensions, and up-sample the time dimension to the length of the target segmentation mask (45). Details of the architecture are presented in Appendix B, available in the online supplemental material.

1) Generating Train and Test Samples From Laughter Annotations: Given that the examples seen by laughter annotators contained a significant amount of context, using the complete 7 s candidates for the machine learning tasks would not be ideal given the much shorter average duration of laughter.

Fig. 5. Illustration of the process used to select positive laughter samples for our machine learning tasks. Given the binary laughter/non-laughter annotations for a particular segment, we select a location for the window center from the positively-annotated intervals in the signal. We then extract a window of 1.5 seconds around the chosen center. We pad if necessary.

Furthermore, our models made use of fixed size inputs, and the examples rated by annotators were not fixed length. To address the situation, we used the continuous binary labeling signal as reference, and sampled shorter positive windows around its positive sections (ie. exactly where laughter was detected to have occurred). Fig. 5 shows a simplified depiction of the process. Given a binary annotation signal with at least one positive segment, we consider the intervals within its positive segments as candidate window locations. We sample uniformly from these locations to select the window center, which determines the limits of the window. For negative examples (ie. with no positive segments), we consider every location in the signal to be a candidate for the window center (ie. we perform a random crop).

To determine the size of the window, we looked at the distribution of laughter lengths, as obtained from our continuous annotations. The average laughter length was 1.14 s, with a long-tailed distribution such that 80 percent of laughs were under 1.56 s. We chose a length of 1.5 s as this length guarantees that most laughter segments will be contained in the window without excessive non-laughter context.

In evaluation, to avoid randomness, instead of the sampling procedure the window is always centered on a positive segment for positive examples. For negatives, the window is always in the middle of the complete candidate.

We followed the same process for the three tasks of laughter detection, intensity estimation and segmentation, but the labels differ per task. For detection, the sample is labeled positive when it comes from a positive annotation segment, and negative otherwise. For intensity estimation, the segment is labeled with the intensity label (Likert scale 1–7) for the laughter candidate. Negative samples were included, and assigned an intensity of zero. For segmentation, the target is a vector corresponding to the continuous binary annotations (30 fps) within the target window (vector of size 45 for our 1.5 s windows).

Note that our annotation study involved two raters per candidate and condition. Both of these continuous ratings are included in the sampling process for each epoch.

2) Evaluation Procedure: For evaluation, we made use of standard metrics for each task. For classification, we make use of the area under the ROC curve (AUC), a metric designed for binary classification and invariant to class imbalance. For regression, we make use of Mean Squared Error (MSE). We also make use of AUC for segmentation, where we treat every window element as one separate prediction. Although metrics like Intersection over Union (IoU) are more commonplace in segmentation, we made use of AUC due to it not being affected by class imbalance.

We evaluated via 10-fold cross-validation, to obtain an aggregated performance measure over the whole dataset. We used the first fold for tuning the number of epochs to train for (per combination of modalities) and excluded the first fold from evaluation.

TABLE I
PRECISION, RECALL AND INTER-ANNOTATOR AGREEMENT AND SIMILARITY MEASURES ACROSS MODALITIES

(a) Laughter detection inter-rater agreement (Cohen's Kappa)

Condition	Audio-only	Video-only	Audiovisual
Audio-only	0.823 (0.153)		
Video-only	0.396 (0.186)	0.550 (0.146)	
Audiovisual	0.795 (0.144)	0.424 (0.183)	0.805 (0.144)

(b) Laughter intensity inter-rater agreement (Krippendorff's alpha)

Condition	Audio-only	Video-only	Audiovisual
Audio-only	0.664 (0.162)		
Video-only	0.237 (0.228)	0.394 (0.279)	
Audiovisual	0.663 (0.168)	0.267 (0.239)	0.697 (0.165)

(c) Laughter segmentation inter-rater similarity (mean IoU)

Condition	Audio-only	Video-only	Audiovisual
Audio-only	0.612 (0.341)		
Video-only	0.378 (0.402)	0.522 (0.409)	
Audiovisual	0.578 (0.358)	0.419 (0.406)	0.661 (0.337)

VI. RESULTS

A. Comparison of Human Laughter Annotation Agreement Across Modalities

To test our hypotheses around differences in annotations across modalities, we started by calculating inter-annotator agreement within and across modalities via pairwise computation of agreement metrics (Section V-C). Tables I(a) and (b) show the results of our agreement calculations for laughter detection and intensity rating. Note that within-modality calculations are averages over 24 (pairwise) comparisons and between-modality calculations are averages over 96 pairs. Standard deviations are shown in parentheses (calculated across pairs). Agreement scores for laughter detection Table I(a) show that the audio and audiovisual conditions have greater within and between modality agreement scores (~ 0.8), with video being significantly lower ($0.396 - 0.550$). The video condition had higher within-condition agreement (0.550) than agreement with other modalities ($0.396 - 0.424$).

Agreement in intensity estimation Table I(b) shows a similar trend. The lowest agreements, once again, were found between audio and video (0.237 ± 0.228) and between audiovisual and

TABLE II
PRECISION AND RECALL W.R.T. TO ANNOTATION REFERENCE

	Audio-only	Video-only	Audiovisual
Precision	0.9645	0.8915	0.9812
Recall	0.9405	0.7024	0.9578

video conditions (0.267 ± 0.239). These are lower than all within-modality agreement scores, even that of video. This suggests that the *concept* of laughter intensity was perceived differently when audio was available and when it was not. Note that agreement in laughter intensity was only calculated between examples labeled positively (as laughter) so that scores are not biased by detection ratings. This resulted in the exclusion of 36% of total ratings (from 3954 to 2531).

We tested the effect of annotation condition on intensity ratings via a linear mixed effects model with the condition as fixed effect. The annotator ID was used as grouping variable (random effect) to control for annotator-specific variance. We fitted the model only on the subset of positive laughter annotations. We found the condition to have a significant effect on intensity ($p = 0.00223$). A cluster bootstrap analysis revealed that laughter was annotated as being significantly less intense in audiovisual (95% confidence interval of $[-0.44, -0.0406]$) and video-only conditions (95% CI of $[-0.45, -0.0482]$). This is a relatively small effect considering the scale of our intensity ratings (1–7).

To get further clarity about the quality of video-based annotations, we compared them to *reference annotations* from the audiovisual condition. We consider the audiovisual condition to be the most ideal one due to annotators having access to both modalities. However, laughter is not always a clear signal and therefore we consider this to be a *reference* set rather than ground truth. We derived this set of binary labels via majority voting, for each $(candidate, condition)$ pair, on the annotator ratings (2), and the expert rating (1), for a total of three votes. We used this reference set for calculation of precision and recall scores.

Table II shows the precision and recall scores for the three annotation modalities w.r.t. the reference annotations. Results show that false positives are rare in our annotations. Recall scores show more differences, with video being lower than both audio and audiovisual scores. This aligns with our hypothesis that the video modality is not enough to detect many episodes of laughter (ie. large number of false negatives). As expected, the audiovisual condition had the highest precision and recall. Note however that reference annotation were obtained from audiovisual labels, and this might cause the numbers to be artificially inflated.

Comparing agreement in localization of laughter is less straightforward, since multiple variables are involved. We did so via computing annotation similarity in corresponding candidates using Intersection over Union (Section V-C). Table I(c) shows a behavior similar to that observed in detection and intensity agreement (Table I(a) and (b)), with similarity in the video condition being even closer to audio(visual) conditions. For a qualitative analysis of agreement in segmentation, we

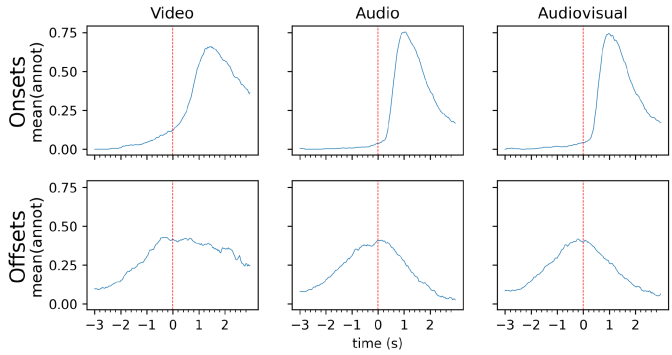


Fig. 6. Aggregated onsets and offsets w.r.t. reference annotations from different modalities.

plotted the mean value of annotations, across different examples, around reference onsets (rising edge of the binary signal) and offsets (falling edge). Ideally, annotators would agree exactly on the onset of the laugh and we would observe a step-like plot. In practice, onsets and offsets vary per annotation and a curve is observed. Fig. 6 shows the mean value of annotations around onsets (key pressed) and offsets (key released). These are aggregated over different laughter samples, and show once again better agreement when audio is present. Offsets display less agreement (flatter shape) than onsets. Note that due to our intention to capture spontaneous annotations of laughter, we did not instruct annotators about the inclusion of the final inhalation as part of the laugh, and this is a possible reason for less offset agreement. It is also possible that offsets tend to be more gradual than onsets on average, but this has not been verified.

We complete our analysis by looking at annotator confidence, as an indication of the difficulty of the task in each modality. Fig. 7 we plot the distribution of laughter intensity and confidence values for the three conditions. We used a Likert scale for both of these ratings, and the distributions are therefore discrete. While intensity distributions are similar across the three conditions, the confidence histograms make clear how much more challenging the video-only condition was to annotators. The wider distribution reveals a clear correlation between laughter intensity and confidence in their annotation, as would be expected.

1) The Role of Laughter Intensity: The results in Section VI-A showed that video-only laughter annotations have lower recall than audio-only annotations. We hypothesized, however, that this is likely due to the difficulty of detecting low-intensity laughs, which are likely to have less salient associated body movements.

To verify this, we separated our dataset by laughter intensity. We obtained a single consolidated audiovisual intensity rating per example by averaging the intensity ratings from both annotators. We then separated the dataset into 10 intensity buckets, from lowest to highest intensity. To ensure a sufficient number of samples per bucket, we used percentiles to define the bucket sizes, such that bucket i includes laughs between the $(i \times 10)th$ and $((i+1) \times 10)th$ percentiles of intensity. We computed recall for each bucket. Fig. 8 plots the results

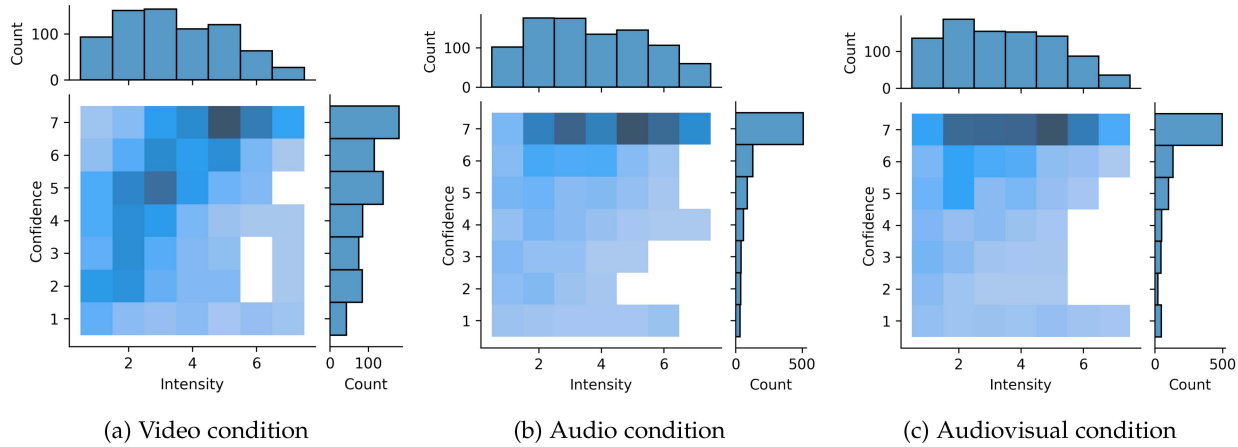


Fig. 7. Joint distribution of confidence and intensity values. Both were annotated using a Likert scale (1–7). Confidence indicates the confidence of the annotator on their laughter annotation for the candidate segment.

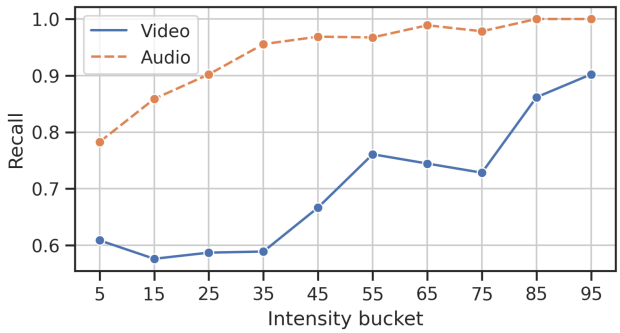


Fig. 8. Laughter recall against (audiovisual) intensity of the laughs. The x axis indicates the middle of percentile bucket (eg. 15 is the bucket with laughs between the 10th and 20th percentile). As intensity increases the recall of video-only annotations approaches that of audio-only annotations.

of this analysis. As expected, recall of both audio and video conditions increases with the audiovisual intensity of the laugh. As hypothesized, video recall tends to approach audio recall for the most intense laughs. It stands out, however, that the gap between them never closes completely, even for the 10% most intense laughs. This can be understood in the light of the findings of Section VI-A, were it was shown that intensity ratings in the audio and audiovisual have high agreement, but they both have low agreement with the video-only ratings. Our consolidated audiovisual intensity ratings, therefore, do not reflect intensity as perceived in the video-only condition. These results align with previous work which observed a positive relationship between laughter intensity and automatic classification performance [48], suggesting that it is not unique to human raters.

B. *Effect of Labeling Modality on Supervised Laughter Tasks*

Although the analysis of inter-annotator agreement performed in the previous section is relevant to understanding differences in labels themselves, it does not ultimately answer the question of how useful annotations acquired from different modalities are for training automated models.

The answer to this question is nuanced. We might have access to video-based annotations of laughter, and want to understand if training a video-based action recognition model with them would help detect vocalizations of laughter. However, asking the reverse question is also of interest: would audio-based annotations result in a model capable of detecting the characteristic body movements of laughter? Furthermore, would audio-based annotations be the most appropriate, or would it be preferable to label the same modality that is input to the model?

The goal of this section is to investigate the impact of annotation modality on trained model performance. Machine learning methods can naturally accept different modalities of input data and we are interested in the relationship and possible interactions between input modality, training label modality, and testing label modality.

To this end, in line with the tasks that annotators performed in our human study, we trained and evaluated models for the tasks of laughter detection, intensity estimation and segmentation (Section V-D). For each of these tasks, we evaluated models for all possible combinations of six different input types (acceleration, audio-only, video-only, video+acceleration, audio+video, audiovisual), training label modalities (audio, video, audiovisual) and testing label modalities (audio, video, audiovisual). We used acceleration as an additional input to leverage the wearable data available in our dataset. Wearable acceleration has been found in previous work to be a useful proxy for body movement. Positive and negative examples were generated for our experiments from the human laughter annotations per the procedure in Section V-D1. We evaluated each model using 10-fold cross-validation and the Area under the ROC Curve (AUC) as evaluation metric, as explained in section V-D2.

Fig. 9 presents the results of our machine learning runs. For readers' convenience we may refer to the results in the tables using the abbreviations in the column labels. For example, $I = Acceleration, Tr = Video, Te = Video$ localizes the fourth cell in the *Acceleration* column.

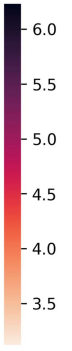
It is clear that for all tasks (audio-)visual inputs trained (audio-)visual labels ($I = Audio|Audio + Video$,

(a) Classification into laughter / non-laughter (AUC, higher is better).

Train label (Tr)	Test label (Te)	Model input (I): Video	Audio	Audio+Video	Model input, acceleration methods (I): Acceleration	Accel+Video	Accel+Audio+Video
Audio	Audio	0.708 (0.069)	0.904 (0.017)	0.897 (0.025)	0.704 (0.076)	0.756 (0.080)	0.901 (0.021)
	Video	0.658 (0.043)	0.609 (0.031)	0.644 (0.027)	0.602 (0.084)	0.659 (0.043)	0.642 (0.017)
	Audiovisual	0.704 (0.093)	0.878 (0.020)	0.874 (0.041)	0.673 (0.063)	0.745 (0.083)	0.875 (0.032)
Video	Audio	0.680 (0.073)	0.782 (0.030)	0.780 (0.075)	0.662 (0.043)	0.728 (0.074)	0.778 (0.065)
	Video	0.707 (0.036)	0.593 (0.044)	0.711 (0.030)	0.640 (0.036)	0.715 (0.031)	0.730 (0.021)
	Audiovisual	0.696 (0.048)	0.777 (0.036)	0.790 (0.051)	0.635 (0.034)	0.737 (0.047)	0.791 (0.042)
Audiovisual	Audio	0.702 (0.073)	0.904 (0.014)	0.889 (0.020)	0.697 (0.068)	0.745 (0.071)	0.893 (0.026)
	Video	0.673 (0.027)	0.609 (0.023)	0.659 (0.020)	0.607 (0.054)	0.679 (0.040)	0.663 (0.030)
	Audiovisual	0.699 (0.089)	0.882 (0.021)	0.872 (0.042)	0.667 (0.057)	0.738 (0.079)	0.876 (0.037)

(b) Regression of laughter intensity (MSE, lower is better).

Train label (Tr)	Test label (Te)	Model input (I): Video	Audio	Audio+Video	Model input, acceleration methods (I): Acceleration	Accel+Video	Accel+Audio+Video
Audio	Audio	5.25 (0.73)	3.24 (0.52)	3.86 (0.43)	5.00 (0.60)	5.20 (0.98)	3.92 (0.44)
	Video	5.37 (1.12)	5.80 (1.07)	6.23 (1.22)	4.85 (0.47)	5.49 (1.33)	6.22 (1.23)
	Audiovisual	4.82 (0.92)	3.15 (0.53)	3.87 (0.65)	4.59 (0.50)	4.77 (0.89)	3.90 (0.65)
Video	Audio	5.67 (0.48)	4.53 (0.64)	5.06 (0.86)	5.53 (0.72)	5.40 (0.62)	5.07 (0.71)
	Video	4.04 (0.44)	4.78 (0.62)	4.70 (1.33)	4.25 (0.45)	4.08 (0.55)	4.53 (1.19)
	Audiovisual	4.89 (0.73)	4.10 (0.66)	4.64 (1.29)	4.87 (0.76)	4.80 (0.71)	4.55 (1.10)
Audiovisual	Audio	5.12 (0.71)	3.31 (0.40)	4.04 (1.14)	5.11 (0.61)	5.16 (0.91)	4.13 (1.29)
	Video	5.00 (0.74)	5.30 (0.61)	5.58 (1.07)	4.50 (0.37)	5.11 (0.81)	5.59 (1.02)
	Audiovisual	4.62 (0.58)	3.12 (0.44)	3.92 (1.14)	4.59 (0.51)	4.79 (0.78)	4.05 (1.21)



(c) Laughter segmentation (AUC, higher is better).

Train label (Tr)	Test label (Te)	Model input (I): Video	Audio	Audio+Video	Model input, acceleration methods (I): Acceleration	Accel+Video	Accel+Audio+Video
Audio	Audio	0.700 (0.029)	0.806 (0.024)	0.822 (0.024)	0.729 (0.025)	0.737 (0.019)	0.820 (0.015)
	Video	0.668 (0.020)	0.626 (0.036)	0.683 (0.025)	0.686 (0.038)	0.689 (0.025)	0.692 (0.033)
	Audiovisual	0.708 (0.020)	0.787 (0.038)	0.809 (0.026)	0.715 (0.021)	0.737 (0.012)	0.808 (0.023)
Video	Audio	0.686 (0.025)	0.758 (0.024)	0.775 (0.019)	0.652 (0.026)	0.710 (0.025)	0.779 (0.017)
	Video	0.704 (0.025)	0.609 (0.034)	0.714 (0.026)	0.662 (0.026)	0.730 (0.013)	0.731 (0.023)
	Audiovisual	0.697 (0.026)	0.745 (0.034)	0.774 (0.019)	0.658 (0.023)	0.724 (0.022)	0.782 (0.007)
Audiovisual	Audio	0.713 (0.031)	0.804 (0.023)	0.819 (0.026)	0.713 (0.025)	0.736 (0.027)	0.819 (0.020)
	Video	0.681 (0.017)	0.622 (0.029)	0.689 (0.018)	0.676 (0.030)	0.698 (0.029)	0.702 (0.023)
	Audiovisual	0.721 (0.020)	0.788 (0.032)	0.814 (0.029)	0.707 (0.032)	0.737 (0.018)	0.811 (0.023)

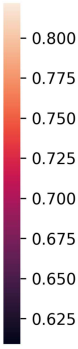


Fig. 9. Results of our machine learning experiments (10-fold cross-validation). Columns correspond to different model input modalities. Rows correspond to training label modality and testing label modality. For example, $Audio > Video$ indicates a model trained with labels acquired from audio alone, and tested on labels acquired from video alone.

$Tr = Audio|Audiovisual$) had the best performances, except when applied to video-based labels ($Te = Video$). This is likely explained by these methods detecting many positives that are not labeled in video, due to having low body movement intensity. In defense of video-based labeling, it stands out that models with video inputs show no significant differences in performance across training and testing label modalities ($I = Video$). In other words, the modality used for labeling had no effect on the final performance of video models. The acceleration, and video+acceleration methods had a similar behavior, with no

significant differences due to training label modality. This provides some support for the use of video labelling for model inputs capturing body movement information. Furthermore, video labels were enough for training an audio-based detection method with an AUC of 0.782 ($I = Audio$, $Tr = Video$, $Te = Audio$), a performance drop of less than 0.15 AUC with respect to audio labels.

Note that classification results Fig. 9(a) display a pattern similar to that of segmentation (Fig. 9(c)). For segmentation, however, scores of audio-based methods ($I = Audio$) are lower than for classification, while the scores of video-based methods $I = Video$) remain the same, making the video-based methods more competitive with the audio-based ones, though still significantly worse-performing for most label combinations. Regarding the acceleration modality, it stands out that $Acceleration + Video$ methods often improved over both modalities in isolation, supporting the idea that these modalities are complementary.

The results of intensity regression methods Fig. 9(b) are more particular. In contrast to classification and segmentation, most multimodal models performed worse (higher MSE) than audio-only models for the same labels (ie. $I = Audio$ generally has the lowest MSE), meaning that adding input modalities tended to affect the model. We also observe that video and acceleration regression models perform best when trained and tested on video labels, but training on audio and testing on video or vice-versa results in some of the worst performances. This aligns with the findings from the annotation experiments that intensity of laughter in the video and audio modalities are incongruent.

VII. Discussion

Our inter-rater agreement results present evidence that annotation of laughter occurrence, intensity and temporal extent can differ substantially across annotation modalities. Per our hypothesis, video annotations had lower agreement than audio and audiovisual ones. When comparing against audiovisual reference annotations, we found recall to be worse in the video condition. Differences in precision scores were lower, with all modalities being close to the 90% to 95% range. These findings suggest that video-based annotation of laughter, while feasible, should not be used in applications requiring high recall. Zooming into the issue of low recall revealed that recall improves for video annotations the more intense the laughs being considered, likely as a result of higher saliency of body movement cues. In the light of previous work [41], this means that video-based laughter annotations are more likely to capture humorous laughter, strongly associated to high intensities, than the more common rule-bound conversational laughter.

Regarding differences between audio and audiovisual conditions, our results revealed high within and between-condition agreement (0.8 for detection, 0.66 for intensity estimation) between them. These results validate the use of audio as primary modality for laughter annotation, but they are not without nuance. Although they indicate that there was a more clear shared concept being annotated when audio was present, video annotations had higher within-condition agreement than agreement with audio and audiovisual annotations. This suggests that there is a different concept being perceived in the video condition with some consistency. In other words, there appears to be incongruence in the perception of laughter occurrence across modalities. Given the low recall of the video condition, we interpret this to indicate that false negatives (w.r.t. audiovisual reference) are missed systematically, likely due to the absence or subtlety of their visual cues. Systematic false positives across annotators also likely contribute to these results, though to a smaller degree. Speech laughter too could play a role in this incongruence. Because no specific instructions were given to annotators on whether to label speech laughter as positive or negative, it is possible that speech laughter was less ambiguous in the video condition than when audio was present.

These results set the stage for the question explored in our machine learning analysis: is perception of laughter in the visual modality a meaningful concept to annotate for the purpose of building detectors, despite it's incongruence with audiovisual laughter?

Importantly, we measured a similar incongruence in laughter intensity ratings, where only positively-labeled segments were included in the agreement calculations, indicating that laughter intensity is not perceived in the same way when audio is present and when it is not. Such incongruences in laughter intensity across modalities have only been studied in the context of laughter synthesis. Niewiadomski et al. found that synthetic laughter episodes with incongruent body movement and vocalization intensities were rated as less believable [8]. This would seem to go against our results, which suggest that significant incongruence exists in in-the-wild laughter perception. However, the magnitude of the incongruencies used (which can be controlled in a synthesis study, but not in the wild) could explain this discrepancy.

Our results have implications in studies of laughter intensity [4], [5], [6], [7], [8], [9], [10], suggesting that the concept of laughter intensity should not be treated as a scalar property of the laughter episode, but rather as a nuanced evaluation affected especially by the modalities available to the observer. In particular, the question of whether a clear distinction should be made between the intensity of body movements and the intensity of the sound of laughter deserves consideration. McKeown et al. already asked the question of whether laughter body movement intensity itself should be considered multi-dimensional [6], but the distinction between visual and auditory intensity has not been considered before, to the best of our knowledge.

Our findings lead us to the fundamental question of what is laughter intensity in the wild. Are the observed differences across modalities mainly a product of imperfect recording conditions, or would we observe them too under ideal conditions? (eg. in face-to-face interactions). While in our dataset subjects prioritized audio in the multimodal condition, it is not clear if body movement information would be prioritized in other datasets in which it is easier to perceive (ie. with consistent access to the face or upper body), or in which the audio is harder to perceive. We consider it likely that in such cases visual information will play a more important role. Although the amount of temporal context (Section V-A2) was the same across modalities, the amount of visual context (Section V-A3)

is unique to the video condition and could play a role in our results as annotators use cues from interlocutors to interpret laughter intensity (and also occurrence). More work is necessary to provide an answer to these questions.

Despite the lower inter-annotator agreement in the video condition, our machine learning experiments with different combinations of model inputs, training label modalities, and testing label modalities, revealed that model performance was the same across labels for models trained using video and acceleration inputs, both of which capture body movements. This was regardless of the evaluation modality. In other words, annotating laughter (traditionally understood primarily as a vocalization) from video alone may be perfectly valid when the goal is to optimize model performance. We think that the reason for such results is explained by our human annotation analysis. Concretely, episodes with lower intensity were most commonly missed (w.r.t. to the audiovisual reference). The subtlety of these training samples would presumably make them more challenging for the learning algorithm, and therefore their absence would not have an adverse effect on performance. We obtained these results in a challenging dataset, where many positive (audiovisual) laughter episodes were missed by annotators, and using a modern action recognition 3D-CNN. It is once again unclear whether these results would translate to a dataset with more consistent access to, for example, facial visual information. The presence of visual cues could improve the model, but their subtlety could be a challenge to most state-of-the-art models. More work is warranted in this direction.

Our results provide validation for previous works using video-only labelling to train laughter assessment models from body movements [4], [65], and datasets providing video-only annotations [16]. Recording audio is not only a technical challenge (especially for large groups), but the use of video labeling is also more privacy conscious as it avoids the need for recording the content of conversations. However, the fact that annotations obtained from video are largely incongruent with audiovisual annotations should be a consideration in studies of laughter.

We think that these results could have wider implications if they generalize to other multi-modal social signals with manifestations in body movement. Speaking status (or voice activity) and back-channels have been of interest in previous work [51], [82]. Video-only annotations of speaking status have been used in previous work [16], [17], [19], but the implications in model performance of this annotation choice have not been explored. Our results would suggest that it is possible to annotate speaking from video alone without an adverse effect on the model's ability to detect speech, but further work is necessary to provide validation for other multimodal social signals besides laughter.

A. Limitations

We consider the main limitation of our work to be that we used only one dataset in our experiments. Our dataset is however representative of one of the most challenging scenarios for perception of laughter from video: with little access to the face of the participants, different views and distances to camera, low light conditions, and significant occlusion of parts of the body from other participants in the scene. We therefore considered it a useful data point to study. We expect that more traditional front-facing datasets with consistent access to the body and face of the subjects will result in lower differences in agreement and model performance between the video condition and the audio and audiovisual ones. We think it is possible that clear access to the face will negate the incongruence observed in laughter intensity ratings, since facial features may share more information with the laughter vocalization than overall body movement does. Finally, obtaining annotations of speech-laughter may have allowed us to better understand the role of speech laughter in the incongruence between audio and video based annotations.

Acknowledgments

We acknowledge our crowd-sourced annotators who very conscientiously participated in our preliminary tests and final study, and in some cases provided valuable voluntary feedback.

References

[1] C. Darwin, "Expression of the emotions in man and animals," *Nature*, vol. 36, no. 926, pp. 294–295, 1887, ISBN: 0195158067.
[2] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, *Social Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.
[4] M. Mancini, G. Varni, D. Glowinski, and G. Volpe, "Computing and evaluating the body laughter index," in *Proc. Int. Workshop Hum. Behav. Understanding*, 2012, pp. 90–98, ISBN: 9783642340130.
[5] R. Niewiadomski, J. Urbain, C. Pelachaud, and T. Dutoit, "Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases," in *Proc. 4th Int. Workshop Corpora Res. Emotion*, 2012, pp. 25–32.
[6] G. McKeown et al., "Human perception of laughter from context-free whole body motion dynamic stimuli," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 306–311.
[7] E. Di Lascio, S. Gashi, and S. Santini, "Laughter recognition using non-invasive wearable devices," in *Proc. 13th EAI Int. Conf. Pervasive Comput. Technol. Healthcare*, 2019, pp. 262–271.
[8] R. Niewiadomski, Y. Ding, M. Mancini, C. Pelachaud, G. Volpe, and A. Camurri, "Perception of intensity incongruence in synthesized multimodal expressions of laughter," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 684–690, ISBN: 9781479999538.
[9] K. El Haddad, S. N. Chakravarthula, and J. Kennedy, "Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles," in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 259–263, doi: 10.1145/3340555.3353764.
[10] W. Curran, G. J. McKeown, M. Rychlowska, E. André, J. Wagner, and F. Lingenfelser, "Social context disambiguates the interpretation of laughter," *Front. Psychol.*, vol. 8, pp. 1–12, 2018.
[11] C. Mazzocconi, Y. Tian, and J. Ginzburg, "What's your laughter doing there? A taxonomy of the pragmatic functions of laughter," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1302–1321, Third Quarter 2020.
[12] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 5117–5120.
[13] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image Vis. Comput.*, vol. 31, no. 2, pp. 186–202, 2013.
[14] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Commun.*, vol. 49, no. 2, pp. 144–158, 2007, ISBN: 0167–6393.
[15] K. P. Truong and J. Trouvain, "On the acoustics of overlapping laughter in conversational speech," in *Proc. Interspeech*, 2012, pp. 850–853.
[16] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during

free-standing conversations and speed dates," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 113–130, First Quarter 2021.
[17] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Pers. Ubiquitous Comput.*, vol. 21, no. 4, pp. 723–737, 2017.
[18] J. Vargas and H. Hung, "CNNs and fisher vectors for no-audio multimodal speech detection," in *Proc. Work. Notes Proc. MediaEval 2019 Workshop*, 2019, pp. 11–13.
[19] C. Raman, J. Vargas-Quiros, S. Tan, E. Gedik, A. Islam, and H. Hung, "ConfLab: A rich multimodal multisensor dataset of free-standing social interactions in the wild," Jul. 2022, *arXiv:2205.05177*.
[20] J. Ginzburg, E. Breitholtz, R. Cooper, J. Hough, and T. Ye, "Understanding laughter," in *Proc. 20th Amsterdam Colloq.*, 2015, Art. no. 11.
[21] K. Oatley and P. Johnson-Laird, "Cognitive approaches to emotions," *Trends Cogn. Sci.*, vol. 18, no. 3, pp. 134–140, Mar. 2014.
[22] K. R. Scherer, "The dynamic architecture of emotion: Evidence for the component process model," *Cogn. Emotion*, vol. 23, no. 7, pp. 1307–1351, Nov. 2009.
[23] P. Glenn, *Laughter in Interact.*. Cambridge, U.K: Cambridge Univ. Press, 2003. [Online]. Available: https://www.cambridge.org/core/books/laughter-in-interaction/4629463A15293CFEBD21EE70AAC966F2
[24] P. Glenn and E. Holt, *Studies of Laughter in Interact.*, P. Glenn and E. Holt Eds., London, U.K.: Continuum Press, May 2013. [Online]. Available: http://www.bloomsbury.com/UK/studies-of-laughter-in-interaction-9781441164797/
[25] G. Jefferson, "On the organization of laughter in talk about troubles," in *Structures of Social Action*, J. M. Atkinson Ed., Cambridge, U.K.: Cambridge Univ. Press, 1985, pp. 346–369. [Online]. Available: https://www.cambridge.org/core/books/structures-of-social-action/on-the-organization-of-laughter-in-talk-about-troubles/7BA34066E5BA65A405570ECA8418B688
[26] M. Gervais and D. S. Wilson, "The evolution and functions of laughter and humor: A synthetic approach," *Quart. Rev. Biol.*, vol. 80, no. 4, pp. 395–430, 2005.
[27] R. I. M. Dunbar, "Laughter and its role in the evolution of human social bonding," *Philos. Trans. Roy. Soc. B: Biol. Sci.*, vol. 377, no. 1863, Sep. 2022, Art. no. 20210176. [Online]. Available: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2021.0176
[28] M. Miller and W. F. Fry, "The effect of mirthful laughter on the human cardiovascular system," *Med. Hypotheses*, vol. 73, no. 5, pp. 636–639, Nov. 2009.
[29] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust laughter detection in noisy environments," in *Proc. Interspeech*, 2021, pp. 2481–2485.
[30] S. Petridis, M. Leveque, and M. Pantic, "Audiovisual detection of laughter in human-machine interaction," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 129–134.
[31] H. J. Griffin et al., "Laughter type recognition from whole body motion," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 349–355, ISSN: 2156–8111.
[32] H. J. Griffin et al., "Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 165–178, Second Quarter 2015.
[33] R. Niewiadomski and C. Pelachaud, "Towards multimodal expression of laughter," in *Proc. Int. Conf. Intell. Virtual Agents*, 2012, Art. no. 6221.
[34] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André, "The belfast storytelling database: A spontaneous social interaction database with laughter focused annotation," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 166–172.
[35] M.-P. Jansen, K. P. Truong, D. S. Nazareth, and D. K. J. Heylen, "Introducing MULAI: A multimodal database of laughter during dyadic interactions," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 4333–4342.
[36] R. Provine, *Laughter: A Scientific Investigation*. London, U.K.: Penguin Press, 2001.
[37] E. J. Capistrano, K. A. R. Espiritu, M. Tandoc, J. K. G. Lim, and J. Cu, "Classifying laughter using the component process model," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact.*, 2022, pp. 1–5, ISSN: .
[38] R. R. Provine, "Laughter punctuates speech: Linguistic, social and gender contexts of laughter," *Ethology*, vol. 95, no. 4, pp. 291–298, 1993.
[39] K. P. Truong, J. Trouvain, and M.-P. Jansen, "Towards an annotation scheme for complex laughter in speech corpora," in *Proc. Interspeech*, 2019, pp. 529–533.
[40] K. El Haddad, H. Cakmak, and T. Dutoit, "On laughter intensity level: Analysis and estimation," in *Proc. Laughter Workshop*, 2018, pp. 34–39.
[41] S. Dupont et al., "Laughter Research: A Review of the ILHAIRE Project," in *Toward Robotic Socially Believable Behaving Systems*. Berlin, Germany: Springer, 2016.
[42] E. Holt, "The last laugh: Shared laughter and topic termination," *J. Pragmatics*, vol. 42, no. 6, pp. 1513–1525, 2010.
[43] N. O'donnell-Trujillo and K. Adams, "Heheh in conversation: Some coordinating accomplishments of laughter," *Western J. Speech Commun.*, vol. 47, no. 2, pp. 175–191, 1983.
[44] A. Wood and P. Niedenthal, "Developing a social functional account of laughter," *Social Pers. Psychol. Compass*, vol. 12, no. 4, 2018, Art. no. e12383.
[45] K. Truong and D. V. Leeuwen, "Automatic detection of laughter," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 485–488, ISBN: 1855212986.
[46] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *Proc. Belgian/Netherlands Artif. Intell. Conf.*, 2008, pp. 351–352.
[47] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *Why When Vis.*, vol. 13, no. 2, pp. 216–234, 2011.
[48] H. Bohy, K. El Haddad, and T. Dutoit, "A new perspective on smiling and laughter detection: Intensity levels matter," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact.*, 2022, pp. 1–8, ISSN: .
[49] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea: IEEE, 2019, pp. 6201–6210. [Online]. Available: https://ieeexplore.ieee.org/document/9008780/
[50] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri, "Automated laughter detection from full-body movements," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 1, pp. 113–123, Feb. 2016.
[51] C. Beyan, M. Shahid, and V. Murino, "RealVAD: A real-world dataset and a method for voice activity detection by body motion analysis," *IEEE Trans. Multimedia*, vol. 9210, pp. 2071–2085, 2020.
[52] L. Cabrera-Quiros, D. M. J. Tax, and H. Hung, "Gestures in-the-wild: Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 138–147, Jan. 2020.
[53] X. Wang, J. Zhu, and O. Scharenborg, "Multimodal fusion of body movement signals for no-audio speech detection," in *Proc. Work. Notes Proc. MediaEval 2020 Workshop*, 2020, Art. no. 3.
[54] K. E. Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laughs: An HMM-based approach for amused speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4939–4943, ISSN: .
[55] S. Petridis, "A short introduction to laughter," 2015. [Online]. Available: https://ibug.doc.ic.ac.UK/media/uploads/documents/shortintrotolaughter.pdf
[56] J. Trouvain, "Phonetic aspects of "speech-laughs,"" 2001. [Online]. Available: https://www.semanticscholar.org/paper/Phonetic-Aspects-of-%22Speech-Laughs%22-Trouvain/5d155268ebe7a565472800c20852a77001c1f8fd
[57] J. Trouvain, "Segmenting phonetic units in laughter," in *Proc. Int. Congr. Phonetic Sci.*, 2003, pp. 2793–2796, ISBN: 1876346485.
[58] H. Bohy, A. Hammoudeh, A. Maiorca, S. Dupont, and T. Dutoit, "Analysis of co-laughter gesture relationship on RGB videos in dyadic conversation context," in *Proc. Workshop Smiling Laughter Across Contexts Life-Span, 13th Lang. Resour. Eval. Conf.*, Marseille, France:European Language Resources Association, 2022, pp. 21–25. [Online]. Available: https://aclanthology.org/2022.smila-1.5
[59] A. Hammoudeh, A. Maiorca, S. Dupont, and T. Dutoit, "Are there any body-movement differences between women and men when they laugh?," in *Proc. Workshop Smiling Laughter Across Contexts Life-Span, 13th Lang. Resour. Eval. Conf.*, Marseille, France:European Language Resources Association, 2022, pp. 30–31. [Online]. Available: https://aclanthology.org/2022.smila-1.9
[60] T. R. Jordan and L. Abedipour, "The importance of laughing in your face: Influences of visual laughter on auditory laughter perception," *Perception*, vol. 39, no. 9, pp. 1283–1285, 2010.
[61] J. Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus," *Lang. Resour. Eval.*, vol. 41, no. 2, pp. 181–190, 2007.

[62] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Mar. 2012, ISBN: 1949–3045.
[63] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2008, pp. 137–148.
[64] M. Mancini et al., "Towards automated full body detection of laughter driven by human expert annotation," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 757–762.
[65] J. Cu, M. B. Luz, M. Nocum, and T. J. Purganan, "Affective laughter expressions from body movements," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2016, pp. 139–145.
[66] Z. H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Comput. Speech Lang.*, vol. 59, pp. 1–21, 2020.
[67] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *J. Acoustical Soc. Amer.*, vol. 88, no. 1, pp. 101–112, Jul. 1990.
[68] X. Alameda-Pineda et al., "SALSA: A novel dataset for multimodal group behavior analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.
[69] M. P. I. for Psycholinguistics, "ELAN [Computer software],"2021. [Online]. Available: https://archive.mpi.nl/tla/elan
[70] J. V. Quiros, S. Tan, C. Raman, L. Cabrera-Quiros, and H. Hung, "Covfee: An extensible web framework for continuous-time annotation of human behavior," in *Understanding Social Behavior in Dyadic and Small Group Interactions*. Cambridge MA, USA: PMLR, Mar. 2022, pp. 265–293, ISSN: .
[71] , "Prolific," 2014. [Online]. Available: https://www.prolific.co
[72] Z. Huang et al., "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, 2015, pp. 41–48.
[73] S. Khorram, M. McInnis, and E. Mower Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1069–1083, Fourth Quarter 2021.
[74] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Second Quarter 2015.
[75] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Gff*, vol. 82, no. 2, pp. 218–226, 2012.
[76] A. Lücking, S. Ptock, and K. Bergmann, "Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann," in *Proc. 9th Int. Conf. Gesture Sign Lang. Hum.-Comput. Interact. Embodied Commun.*, 2011, pp. 129–138.
[77] Y. Mathet, A. Widlöcher, and J.-P. Métivier, "The unified and holistic method gamma for inter-annotator agreement measure and alignment," *Comput. Linguistics*, vol. 41, no. 3, pp. 437–479, Sep. 2015, Cambridge, MA, USA: MIT Press. [Online]. Available: https://aclanthology.org/J15--3003
[78] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1130–1139. [Online]. Available: https://ieeexplore.ieee.org/document/8578222/
[79] H. Fan et al., "PyTorchVideo: A deep learning library for video understanding," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021.
[80] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 3783–3786.
[81] I. Oguiza, "tsai - a state-of-the-art deep learning library for time series and sequential data," 2022. [Online]. Available: https://github.com/timeseriesAI/tsai
[82] K. P. Truong, R. Poppe, I. De Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 2973–2976, 2011.

Jose David Vargas Quiros is currently working toward the PhD degree with the Socially Perceptive Computing Lab, TU Delft, The Netherlands, since 2018. He is interested in multimodal action recognition and conversation quality assessment in-the-wild, the study of interpersonal adaptation and synchrony, and efficient annotation of in-the-wild data.

Laura Cabrera-Quiros is an Assistant Professor with the Costa Rican Institute of Technology (Instituto Tecnológico de Costa Rica), working in the Electronics Engineering Department. Her research focuses on the use of machine learning and non-invasive technologies (e.g. wearable and embedded devices, cameras, physiological sensors) to understand human behavior, monitor health, and improve people's quality of life.

Catharine Oertel is an Assistant Professor with TU Delft, The Netherlands. She is co-Principal investigator of the Designing Intelligence Lab (DI_Lab), an effort aiming to bridge research done in computer science with industrial design engineering. Her research interest includes lies on understanding and modeling human interaction to build socially aware conversational agents able to engage with people in a human-like manner.

Hayley Hung received the PhD degree in computer vision from the Queen Mary University of London, in 2007. She is an Associate Professor with the Socially Perceptive Computing Lab, TU Delft, The Netherlands, where she works since 2013. Between 2010-2013 she held a Marie Curie Fellowship with the Intelligent Systems Lab, University of Amsterdam. Between 2007-2010 she was a postdoctoral researcher with IDIAP Research Institute in Switzerland. Her research interests include social computing, social signal processing, computer vision, and machine learning.