Automatic Recognition of Affective Laughter in Spontaneous Dyadic Interactions from Audiovisual Signals

Reshmashree B. Kantharaju*
ISIR, Sorbonne Université
Paris, France
bangalore_kantharaju@isir.upmc.fr

Fabien Ringeval
LIG, Univ. Grenoble Alpes, CNRS,
Grenoble INP
Saint-Martin-d'Hères, France
fabien.ringeval@imag.fr

Laurent Besacier
LIG, Univ. Grenoble Alpes, CNRS,
Grenoble INP
Saint-Martin-d'Hères, France
laurent.besacier@imag.fr

ABSTRACT

Laughter is a highly spontaneous behavior that frequently occurs during social interactions. It serves as an expressive-communicative social signal which conveys a large spectrum of affect display. Even though many studies have been performed on the automatic recognition of laughter – or emotion – from audiovisual signals, very little is known about the automatic recognition of emotion conveyed by laughter. In this contribution, we provide insights on emotional laughter by extensive evaluations carried out on a corpus of dyadic spontaneous interactions, annotated with dimensional labels of emotion (arousal and valence). We evaluate, by automatic recognition experiments and correlation based analysis, how different categories of laughter, such as *unvoiced laughter*, *voiced laughter*, *speech laughter*, and *speech* (non-laughter) can be differentiated from audiovisual features, and to which extent they might convey different emotions. Results show that voiced laughter performed best in the automatic recognition of arousal and valence for both audio and visual features. The context of production is further analysed and results show that, acted and spontaneous expressions of laughter produced by a same person can be differentiated from audiovisual signals, and multilingual induced expressions can be differentiated from those produced during interactions.

CCS CONCEPTS

• **Information systems → Multimedia databases**; *Presentation of retrieval results*;

KEYWORDS

Laughter, Affective Computing, Context Recognition

ACM Reference Format:
Reshmashree B. Kantharaju, Fabien Ringeval, and Laurent Besacier. 2018. Automatic Recognition of Affective Laughter in Spontaneous Dyadic Interactions from Audiovisual Signals. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3243012

1 INTRODUCTION

Laughter is a non-verbal signal that plays a prominent role in many social situations [11], especially in social interactions where it is about 30 times more frequent than in solitary situations [39]. In general, laughter is considered as a universal social signal which is beneficial for health, as it has been shown to help in reducing stress [5], anxiety, pain and discomfort [28] and also improve the mood [31]. It generally conveys an essential form of social relief during interactions, which might indicate a cooperative intent.

Laughter is a subjective feeling whose expressions might be related to the situation it is produced in. It can occur naturally during friendly social interactions or it can be induced by humorous content like video clips, stories, jokes, games, even though a person is more likely to laugh when in a group or a gathering of closed persons [20]. Spontaneous laughter follows an impulse, which is an urge to laugh without restrain or control over its expression [44]. Although the acoustic pattern is similar to that of natural laughter, acted laughter has been reported to lack emotional content [44].

Several taxonomies exist for laughter, such as those based on the context of production (e. g., spontaneous, acted) [36, 56], or on articulatory properties (e. g., voiced, unvoiced, speech-laugh) [2, 25], or the conveyed emotions (e. g., happiness, hurtful, embarrassment, excitement etc.) [13, 50, 51]. Laughter can be defined – from the production side – as a rhythmic, vocalised and involuntary action caused by the body under certain conditions. It is usually accompanied by smile [9], and involves certain facial and body movements and change in postures [44]. Regarding the timing of a laughter event, an entire occurrence of laughter is usually termed as a *laughter episode* [39, 44]. It includes the vocal and other correlated elements and it is usually made of one or several *bouts*, i. e., a behavioral-acoustic event, including the respiratory, vocal, and facial and skeleto-muscular elements [44].

Contextualisation of laughter with respect to the expressed emotion is important, because laughter can be produced in a large variety of context, and therefore convey specific non-verbal messages. Whereas, Darwin assumed that laughter is an expression of happiness [12], Ekman suggested that it can rely on a combination of several emotions which might differ in their acoustical structure [14]. A review of literature shows that laughter can actually express various emotions like, joy, amusement, surprise, but also nervousness, taunt, embarrassment, contempt [46], sadness [49], or even '*schadenfreude*', i. e., the pleasure in other's misfortune [10, 50, 51].

Automatic recognition of laughter can be useful for multimedia tagging and retrieval i. e., extraction of humorous content or automatically identifying meaningful events in meetings such as topic change or jokes [36]. Another area of application is computer-aided

*Previously affiliated with LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, 700 Avenue Centrale, Saint-Martin-d'Hères, France, 38401

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243012

Table 1: Overview of publicly available audiovisual databases of laughter; A: audio; V: video; P: physiological; K: kinect; FT: face tracking; BM: body movement; * : part of the ILHAIRE Laughter database.

Database	Modality	Type	Participants	Duration	Laughter Instances	Emotion Annotation
AVLC [56]	A, V, FT	Induced, Posed	24	64m 28s	1066	No
BELFAST* [29]	A, V, K	Interactive	21	106m	2336	Yes
BINED* [48]	A, V	Induced	256	29m 45s	289	Yes
MAHNOB [36]	A, V, P	Induced, Posed	22	18m 59s	563	No
MMLI [32]	A, V, P, FT, BM	Induced, Interactive	16	31m	439	No
RECOLA [43]	A, V, P	Interactive	46	17m 58s	974	Yes
SEMAINE* [30]	A, V	Interactive	150	2m 05s	443	Yes

psychotherapy, where a computer can monitor the reactions of the patients through multimodal signals [22]. Moreover, contextual information on affect is necessary to understand the interplay between the kind of laughter and emotion in the enactment. Whereas several studies have reported on the spectrum of emotions that laughter conveys [2, 13, 51], there has been no investigations – to the best of our knowledge – on the automatic recognition of the emotions conveyed by natural expressions of laughter produced during spontaneous interactions.

In order to investigate automatic recognition of emotional laughter from spontaneous multimodal data, we performed annotation of different types of laughter on the RECOLA dataset [43]. A total of 974 instances has been annotated, and will be made publicly available[1] to the community. Further, we conduct experiments to distinguish different categories of laughter, quantify their predictive power for emotion recognition, and the impact of the context. Interplay between categories of laughter, emotional dimensions, and audiovisual descriptors are first investigated with correlation based measures, before reporting results of automatic recognition experiments.

In the remainder of this paper, we present related work on existing audiovisual databases of laughter, including automatic recognition experiments (Sec. 2), then introduce the methodology followed to perform annotation of laughter on the RECOLA database (Sec. 3), and report results on the experiments on the automatic recognition of categories of laughter, their conveyed emotions, and on the context of production (Sec. 4), before concluding (Sec. 5).

2 RELATED WORK

There exists many databases which include annotations of laughter from audiovisual data, but very few are publicly available, and most of them are not specifically dedicated to laughter. In this section, we provide an overview of publicly available audiovisual databases designed for the analysis of laughter. The selected databases provide data across several modalities, languages and methods used to elicit laughter, and have been used widely for analysis and synthesis of laughter. Table 1 provides a general overview of the selected databases and a short description is given for each dataset below.

The AVLC database provides audiovisual recordings of induced and posed laughter elicited from 24 subjects, while watching a funny 10 minutes stitched clip of short videos [56]. A hierarchical protocol was used for the annotation of laughter where each segment was first given a label and sub-labels were added based on the temporal structure or acoustic contents to provide further details on different categories of laughter. Due to the absence of social interactions, the amount of speech laughter and speech data is minimal.

Belfast storytelling database provides naturalistic multimodal data associated with social interactions in a semi-structured storytelling environment [29]. It consists of 21 participants telling stories in groups of three or four in English or Spanish which occasionally led to an open discussion, which facilitated the occurrence of conversational laughter. The laughter segmentation is done on two levels: auditory, and visual cue. Automatic pre-annotation of laughter was done by training a Support Vector Machine on acoustic and visual features [27], which was later refined manually.

The BINED database has 3 sets of audiovisual recordings of emotion elicited from watching video clip (e. g., amusement) and actively engaging participants in series of tasks to induce emotions (e. g., fear, disgust, surprise, frustration) [48]. The first set has been included in the ILHAIRE database and it consists of 565 clips of 113 participants from which 289 instances of laughter were extracted.

MMLI is a multimodal database of laughter with full body movements, facial tracking, audiovisual and physiological data [32]. The participants were asked to perform a set of tasks in groups to record spontaneous as well as controlled laughter. The database is annotated as, Laughter event (time interval in which at least one of the participants laughs) and Laughter episode (single laugh generated by one participant); one laughter event can thus be composed of several laughter episodes.

The SEMAINE database consists of audiovisual recordings of users interacting with limited agents, which present different personalities [30]. Laughter was annotated from the Solid Sensitive Artificial Listening (SAL) part where one participant took the role of the user and another took the role of one of the four SAL characters. In total 443 instances have been annotated from 345 clips from 28 participants. The SEMAINE-SAL database includes time- and value-continuous annotations of emotional dimensions (arousal and valence), but the annotators are not consistent over the recordings.

Experiments on the automatic recognition of laughter from the audiovisual data contained in the SEMAINE-SAL database have been previously reported on a lower amount of instances [35]. Data were partitioned into speaker independent training, validation and test partitions. Mel-Frequency Cepstrum Coefficients (MFCCs) were extracted from the acoustic data, whereas facial animation parameters, automatically computed by a 3D tracker [34], were

[1] https://diuf.unifr.ch/diva/recola/

exploited as visual features. Online standardisation of both acoustic and visual features was performed before learning a 2-class model to distinguish speech vs. laughter utterances for each modality by time delay neural networks. Decision-level fusion was then operated at the frame level by a linear combination of the *a posteriori* probabilities. The reported performance, which was measured by the Unweighted Average Recall (UAR), i. e., the average of the recall of the two classes in percentage, shows that audio data performed best (69.4%), whereas video data performed better than the chance level (56.6%), and the fusion of the two modalities only slightly improved the performance (69.5%).

The MAHNOB database includes multimodal recordings (audio, video and physiological signals) of 22 subjects from 12 different countries watching series of video clips [36]. The main goal was to elicit laughter, but along with it the participants were also instructed to pose a laugh and smile and then were asked to speak in English as well as in their native language to create a multilingual corpus. In addition to that, laughter episodes were further annotated as voiced or unvoiced using a combination of two approaches: manual labeling by two human annotators and automatic detection of unvoiced frames based on the pitch contour computed with PRAAT [6]. Each episode was then assigned a label based on majority voting. Further details on the data are provided using the following categories of annotation: laughter, speech, speech-laugh, posed smile, posed laughter, laughter + inhalation, speech-laugh + inhalation, posed laughter + inhalation, using the ELAN annotation tool [8].

Experimental evaluations have been reported on the MAHNOB dataset for the automatic discrimination between laughter and speech, and between voiced laughter, unvoiced laughter and speech. Evaluations were carried out with a leave-one-subject-out cross-validation methodology. A feedforward neural network was used to train binary classification models. As evaluation metrics, the F1 score and the Classification Rate (CR) are reported. As found in previous experiments, audio features performed best for the 2-class problem (speech vs. laughter), and visual features provided complementary information (feature-level fusion); obtained performance (CR) was 86.2% for audio, 83.9% for video, and 90.1 for audiovisual. A similar contribution of the audiovisual modalities was observed for the 3-class problem: audio: 79.1%; video: 74.5%; audiovisual: 83.2%.

Even though publicly available databases of laughter provide multimodal, multilingual data, with (partially) rich annotations, there is no ratings of affective behaviour available with respect to laughter, except for the SEMAINE database. Another drawback of most existing databases is the lack of spontaneous laughter from natural interactions. Although Belfast Storytelling database provides naturalistic conversational laughter the annotations are minimal and does not provide any further details on laughter categories. Therefore, for this study, we decided to perform annotation of laughter episodes on an existing database of spontaneous socio-affective behaviors.

3 ANNOTATION OF LAUGHTER

The chosen RECOLA dataset [43] consists of natural and spontaneous interactions and makes investigations on laughter realistic for real-life scenarios. The details of the available data set and annotations performed on it are provided in the following sections. We first briefly introduce the database and then detail the methods we used to perform laughter annotation. The scheme followed to obtain different categories of laughter is similar to MAHNOB [36], as we are convinced that such categories (i. e., voiced, unvoiced, speech laughter) not only convey different kinds of acoustic and facial information, but also different kind of affective behaviours, which will be demonstrated in the experimental evaluations.

3.1 RECOLA database

The REmote COLlaborative and Affective interactions (RECOLA) data set is a multimodal database of spontaneous interactions in French which consists of audio, visual, electro-cardiogram (ECG) and electro-dermal (EDA) data recorded continuously and synchronously [43]. This data set was used in the Audio Visual Emotion recognition Challenge (AVEC) for benchmarking multimodal emotion recognition systems [42, 57]. Spontaneous interactions from 53 participants were recorded while solving a collaborative task: "Winter Survival Task"[21] as dyadic teams. The task first involved individual ranking of 15 items in order of importance for survival and then participants had to discuss their rankings in order to reach a consensus. The recordings which are 9.5 hours long were made in isolated rooms free from external noises and the participants were separated in two rooms and interacted through Skype. This scenario is highly relevant for our study, since many naturalistic reasons to laugh are present, such as laughter from embarrassment, knowing they are being recorded; from hesitation or stress relief while discussing with a stranger (less than 20% of participants of the RECOLA dataset knew their teammate well), from humor or discussions perceived to be funny by the participants – usually while discussing the order of importance of items.

Affective behaviour expressed by the participants was annotated with time- and value-continuous emotional dimensions (arousal and valence) by six French-speaking assistants, for the first five minutes of each recording, and for 46 participants. Only the first five minutes of interactions were annotated since participants spent more time discussing their strategies at the beginning of the task and also to limit the amount of data to be annotated. A web-based annotation tool, ANNEMO was used to perform emotion ratings and the annotations were done separately for each emotional dimension, using a slider with values ranging from -1 to +1 and a step of 0.01, obtained values were then re-sampled at a frame rate of 40ms. In order to obtain a single gold-standard from the pool of ratings, each trace (i. e., a time- and value-continuous emotional annotation) was first assigned a weight according to the agreement of this rating with the five others. The final gold-standard was obtained by simply averaging the traces, after normalisation according to the inter-rater agreement; see [41, 42, 57] for more details on the computation of the gold-standard and statistics on inter-rater agreement.

3.2 Laughter episode

Laughter occurs frequently during interaction and it is relatively difficult to identify exactly where it begins and ends. Some studies consider visual information in addition to the audio information for determining the start and end points i. e., *onset* and *offset* respectively. Segmentation and terminology used to define an instance of laughter is heterogeneous due to its multi-disciplinary nature [53]. Several studies in the past have described different ways of segmenting laughter based on temporal characteristics [1, 44, 53]. In this study, we adopt the terminology proposed in [44]. Laughter unlike speech, disrupts normal breathing and audible inhalations can occur during laughter [9]. An entire occurrence of laughter is termed as a *laughter episode* [39, 44] and can be subdivided further into three parts, ***Onset*** : Pre-vocalization period leading to laughter, usually a strong exhalation; ***Apex*** : Period with vocalization or rhythmic exhalation and composed of laughter syllables; ***Offset*** : Post-vocalization period, usually includes an inhalation.

The inclusion of inhalation at the end of post-vocalization period as a part of laughter has been debatable and is still unclear [53]. In [1] it was not considered as a part of laughter. In the AVLC database [56], it was included as a part of laughter and MAHNOB [36] provides two different sets of labels with inclusion and exclusion of inhalation. In this study, we include the post inhalation as we consider it to be a part of laughter.

3.3 Laughter categories

Based on the articulatory properties, laughter can be differentiated as *voiced laughter* and *unvoiced laughter* [2, 25]. Voiced laughter consists of voicing element (i. e., periodic vibrations of vocal folds) and the excitation is partly quasi-periodic. Unvoiced laughter, on the other hand, does not consist of any voicing element and the excitation is fricative. It is more irregular and does not have a rhythmic structure [55]. Also, it has been shown in several instances that speech and laughter can overlap during conversations [13, 33, 52]. This form of laughter is termed as *speech-laugh*. It occurs simultaneously with articulation [53] and is generally longer than just laughter, with more energy variations than regular speech. In some studies it has been annotated as speech [24, 26] and in some it is said to be an independent laughter category [4, 33, 36, 56]. Considering the findings of these studies, we annotated laughter episodes and categorized them as Unvoiced laughter, Voiced Laughter and Speech laughter.

3.4 Annotation

The initial annotations of laughter was done manually by the first author of this study using the Audacity[2] software for the 53 audio clips from the RECOLA dataset. Audio data was primarily used for annotation of laughter episodes and video data was referred when there was ambiguity. Speech laughter was annotated when both laughter and speech occur simultaneously and it does not include speech before and after (if any) the laughter bouts. The laughter instances (excluding speech laughter) were further categorised as *voiced* or *unvoiced*. In literature, there has been several ways proposed to categorise it. It can be done manually by human annotators [4] or automatically using the pitch contour [23] or a combination of both [36]. In this study, we use the voicing probability and unvoiced frame ratio to automatically categorise voiced and unvoiced laughter.

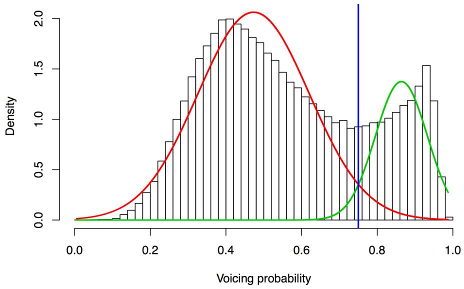

Figure 1: Probability density function of voicing probability to determine the unvoiced ratio threshold value used to distinguish between voiced and unvoiced laughter.

In order to help differentiating *voiced* from *unvoiced laughter*, we computed the voicing probability of each frame using the OPENSMILE[3] acoustic feature extraction toolkit [18]. A histogram of those probabilities was plotted for all laughter episodes and speech episodes in the 53 annotated audio files to observe the distribution. This was done in order to calculate the value of voicing probability threshold which would be used to decide the voiced and unvoiced frames. The data was fitted to two Gaussian distributions to find the mixing ratios and the obtained threshold is 0.76, cf. Figure 1. Based on the voicing probability threshold, for each laughter episode, the unvoiced frame ratio was calculated. For a given laughter episode, all the frames with voicing probability below the threshold value were considered as unvoiced frames and above the threshold as voiced frames. Unvoiced frame ratio for each laughter episode can be calculated simply as the number of unvoiced frames divided by the total number of frames.

Distinguishing unvoiced laughter from voiced laughter was not easy as there is no clear cut-off. The criteria to decide whether a given laughter instance is voiced or unvoiced has been ambiguous. While in [38] a given laughter instance was considered unvoiced if the unvoiced ratio was more than 80%, in [54] laughter was considered voiced if it contained at least one voiced frame. In [36] the cut-off for unvoiced ratio was set at 85% below which the laughter instance was considered as voiced and the episodes were labeled manually by 2 annotators and the final label was assigned based on majority voting. In this study, the unvoiced ratio histogram was plotted for both unvoiced laughter and voiced laughter and we found the right value empirically by looking at segments found at different thresholds. The laughter episodes at different threshold ranges were observed individually. These observations indicated that, a threshold value for unvoiced ratio, lower or higher than 0.75 would mis-classify the voiced laughter and breath laughter episodes. Based on the observations a final threshold value of 0.75 was chosen. The unvoiced frame ratio was calculated based on the voicing probability. If the unvoiced frame ratio was greater than the threshold value the laughter labels were corrected as *Unvoiced Laughter* and if it was less, then the laughter labels were corrected as *Voiced Laughter*.

[2] Version 2.1.2.0 of Audacity(R) recording and editing software.

[3] http://audeering.com/technology/opensmile

Table 2: Number of laughter episodes annotated on the RECOLA database from complete unsegmented and segmented (initial 5 minutes) audio files with emotion ratings.

Type	Unsegmented	Segmented
Unvoiced Laughter (UL)	590	159
Voiced Laughter (VL)	187	62
Speech Laughter (SL)	197	68
All Laughter (AL)	**974**	**289**

RECOLA [43] data as mentioned earlier, is divided into two sets namely, *Unsegmented set* and *Segmented set.* The segmented set provides emotional ratings across the two affect dimensions (i. e., Arousal and Valence) for the initial five minutes. Turn timings are available for the segmented clips to indicate when there is speech vocalisation. Pure speech segments which did not contain any laughter and was at least one second long were extracted. The laughter episodes were labeled as 'VL' for *voiced laughter*, 'UL' for *unvoiced laughter*, 'SL' for *speech laughter* and non laughter speech segments were labeled as 'S'. In total there are 289 instances of laughter and 1619 instances of speech. The unsegmented set only provides annotations of laughter and consists of 974 instances of laughter. Both sets are used for recognition of various kinds of laughter and the segmented set is used for the recognition of emotion conveyed by laughter. Table 2 shows the number of laughter instances annotated in the RECOLA database for the 53 unsegmented (complete audio) and 46 segmented (annotated audio) clips.

Table 3: Z-score of statistics of three acoustic features (voicing probability, pitch, and loudness) for different types of laughter instances computed on segmented audio data using openSMILE; $*$ indicates that there is a statistically significant difference (1-way ANOVA; $p < 0.05$) when compared with those for speech; mean (standard deviation);

Type	VoiceProb	Pitch	Loudness
Unvoiced Laughter	-0.71 (0.82)*	0.38 (1.82)	-0.50 (0.90)*
Voiced Laughter	-0.33 (0.92)*	0.28 (1.53)*	-0.16 (1.21)
Speech Laughter	-0.25 (1.00)*	0.38 (1.37)*	0.14 (1.30)*
All Laughter	-0.47 (0.93)	0.26 (1.58)	-0.20 (1.15)
Speech	0.04 (0.99)	-0.02 (0.95)	0.02 (0.99)

The mean and standard deviation calculated for the laughter and speech episodes from prosodic features (voicing probability, pitch, and loudness) computed on the segmented audio clips, after a z-score normalisation to compensate speaker dependencies, are shown in Table 3. Results show that the voicing probability is significantly (1-way ANOVA) lower for all kinds of laughter instances compared to speech instances, pitch values are higher than speech for all laughter instances (few voicing parts of unvoiced laughter instances might have compromised the pitch estimation algorithm), whereas loudness is lower for unvoiced laughter and voiced laughter, but higher for speech laughter.

Table 4: Statistics on emotional ratings for arousal and valence for the initial 5 minutes of the 46 segmented audio-visual clips for each kind of laughter and speech episode. $*$ indicates that there is a statistically significant difference (1-way ANOVA; $p < 0.05$) with speech instances; mean ± standard deviation.

Type	Arousal	Valence
Unvoiced Laughter	0.08 ± 0.13	0.28 ± 0.12*
Voiced Laughter	0.13 ± 0.12*	0.31 ± 0.13*
Speech Laughter	0.18 ± 0.11*	0.33 ± 0.14*
All Laughter	0.12 ± 0.13	0.30 ± 0.13
Speech	0.10 ± 0.12	0.10 ± 0.12

3.5 Emotional ratings

It has been repeatedly shown that there is a delay in the time-continuous rating of emotional dimensions, and that such delay has an important impact on emotion recognition performance [41, 57]. We therefore included this information when assigning emotion labels to speech and laughter instances. A delay of 2.8 s for arousal and 3.6 s for valence was applied to the gold-standard (values were shifted backward in time); values were taken from the AVEC'16 challenge [57], where machine learning optimisation (grid search) was used in order to estimate those delays. Even though the annotation delay is not constant between raters, sequences, and even over time, we assume that those optimised values for arousal and valence would provide sufficient reliable labels for the corresponding laughter events.

The mean rating of arousal and valence were calculated for each annotated speech/laughter episode. Obtained values are reported in Table 4 and show that there is a statistically significant difference in the emotion conveyed by different kinds of laughter when compared to speech and each other, for both arousal and valence. Moreover, we can clearly see that all laughter instances convey higher valence than speech instances, and unvoiced laughter has the lowest arousal and valence ratings among the laughter categories.

4 EXPERIMENTS

We present in the following sections the methodology that we exploited to perform the automatic recognition of various types of laughter instances, as well as their conveyed emotion. We first introduce the acoustic and video feature sets[4], then present the system and the obtained performance.

4.1 Feature Sets

4.1.1 Audio. In contrast to large scale feature sets, which have been successfully applied to many speech classification tasks [41, 58], smaller, expert-knowledge based feature sets have also shown high robustness for the modeling of emotion from speech [7, 40]. Some recommendations for the definition of a minimalistic acoustic standard parameter set have been recently investigated, and have led to the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and

[4] We did not make use of the physiological signals as those are only available for 27 subjects in total.

Table 5: Statistics (mean and standard-deviation) of the intensity of 17 Facial Action Units (FAUs) estimated from the video recordings with the OpenFace toolkit. $*$ indicates that the difference in the values of given AU for all three kinds of laughter is statistically significant (1-way ANOVA; $p < 0.05$) when compared with those for speech individually.

FAUs		1	2	4	5	6*	7*	9	10*	12*	14*	15	17*	20	23	25*	26	45
Unvoiced Laughter	Mean	0.67	0.40	0.55	0.26	1.98	1.85	0.38	1.15	1.85	1.20	0.73	0.55	0.35	0.47	2.03	1.01	0.79
	StdDev	0.72	0.47	0.53	0.32	0.81	0.90	0.34	0.83	0.77	0.65	0.95	0.52	0.35	0.50	1.14	0.92	0.77
Voiced Laughter	Mean	0.68	0.72	0.49	0.34	1.93	1.94	0.40	1.25	1.77	1.39	0.73	0.59	0.37	0.51	1.95	1.23	1.11
	StdDev	0.96	0.62	0.52	0.40	0.92	0.96	0.32	0.74	0.89	0.94	0.88	0.45	0.35	0.52	1.29	1.13	0.86
Speech Laughter	Mean	0.71	0.41	0.51	0.31	1.88	1.90	0.30	1.05	1.69	1.08	0.97	0.66	0.38	0.48	2.10	1.14	0.83
	StdDev	0.60	0.25	0.44	0.43	0.77	0.79	0.30	0.79	0.87	0.69	0.92	0.51	0.27	0.45	1.23	0.74	0.65
All Laughter	Mean	0.69	0.46	0.53	0.29	1.95	1.88	0.36	1.15	1.80	1.21	0.79	0.59	0.36	0.48	2.03	1.09	0.86
	StdDev	0.72	0.47	0.53	0.32	0.81	0.90	0.34	0.83	0.77	0.65	0.95	0.52	0.35	0.50	1.14	0.92	0.77
Speech	Mean	0.72	0.42	0.53	0.33	0.86	1.34	0.30	0.79	0.87	0.89	0.63	0.85	0.32	0.47	0.92	0.97	0.94
	StdDev	0.64	0.45	0.47	0.35	0.58	0.71	0.32	0.62	0.56	0.68	0.77	0.57	0.30	0.47	0.63	0.76	0.68

to an extended version (eGeMAPS) [17]. The acoustic low-level descriptors (LLD) cover spectral, cepstral, prosodic and voice quality information and are extracted with the openSMILE toolkit [18]. The arithmetic mean and the coefficient of variation are then computed on all LLDs. To pitch and loudness the following functionals are additionally applied: percentiles 20, 50 and 80, the range of percentiles 20 – 80 and the mean and standard deviation of the slope of rising/falling signal parts. Functionals applied to the pitch, jitter, shimmer, and all formant related LLDs, are applied to voiced regions only. Additionally, the average RMS energy is computed and 6 temporal features are included: the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments and the rate of voiced segments per second, approximating the pseudo syllable rate. Overall, the GeMAPS acoustic features set contains 88 features, and its extended version includes 102 features. In addition to those (e)GeMAPS feature sets, we also computed 12 MFCCs and the log-energy, to analyse pure spectral features.

4.1.2 Video. Facial Action Coding System (FACS) is a system used to describe human facial movements by their appearance on the face [16]. Action Units (AUs) are the fundamental actions of individual muscles or groups of muscles. Laughter involves facial movements and is often associated with the emotion of joy and this type of laughter is said to be associated with the Duchenne display[5] [15]. Since, several studies have shown that laughter occurs during various emotional states, we aim to study the intensity of AUs associated with other emotions and not restrict the feature set to the AUs associated with laughter. For this study, we make use of OpenFace[6], an open source tool intended for computer vision and machine learning researchers, to extract facial AUs that will serve as facial descriptors for our study set [3]. The tool offers two kinds of scores for the AU: intensity and presence. The former provides the intensity of 17 AUs on a continuous value scale from 1 (minimally present) to 5 (present at maximum intensity); a score of 0 indicates absence. The latter indicates the presence or absence of 18 AUs (provides an additional AU28, along with those mentioned in Table 5). We exploited the intensity of AUs since we are studying the intensity of affect in laughter. The video feature set consists of the mean intensity and standard deviation values when a AU was present, for the 17 AUs over a given episode. We also calculated the proportion of activation of the AUs by simply dividing the number of frames a given AU was present over the total number of frames for a given episode. In total, the video feature set (FAUs) consisted of 51 features (17x3) associated with facial AUs. Table 5 reports statistics of those AUs for each laughter and speech categories. We observe some AUs e. g., AU6, AU12 that are often associated with laughter are statistically significant when compared with speech.

4.2 System

For the automatic recognition of laughter instances we make use of LIBLINEAR, an open source library for large-scale linear classification [19]. It supports logistic regression and linear support vector machines which has proven to provide state-of-the-art performance on many affect related prediction tasks [42, 57]. A classifier model was built using .mex implementations of LIBLINEAR that interface with MATLAB. The training was optimised with 3 different solvers for classification: L2-regularized L2-loss support vector classification (primal and dual), and L2-regularized L1-loss support vector classification (dual). The complexity parameter was also optimised in a logarithmic scale, with values ranging from 0.0001 to 1. To ensure speaker independent evaluations, we follow a Leave-One-Speaker-Out (LOSO) cross-validation methodology in all the experiments for recognition of laughter and the emotions conveyed by them. In order to compensate speaker dependencies of the features, we investigated three different approaches: (i) *offline-partitioning*, i. e., mean and standard-deviation were computed independently on both training and testing partitions, (ii) *offline-speaker based*, i. e., mean and standard-deviation were computed individually with respect to the speaker, and (iii) *online-partitioning* i. e., mean and

[5] The Duchenne display: joint contraction of facial muscles, i. e., pulling the lip corners backwards and upwards (AU12) and raising the cheeks (AU6) causing eye wrinkles (AU7).

[6] https://github.com/TadasBaltrusaitis/OpenFace

Table 6: Results (%UAR) for 2 class, 3 class and 4 class classification tasks for RECOLA (Segmented and Unsegmented) using different feature sets : 3 audio, 1 video and the combination of best audio (eGeMAPS) and video feature set; best results over the feature sets are highlighted in bold format.

		Segmented set					Unsegmented set				
Type	Case	MFCCs	GeMAPS	eGeMAPS	FAUs	Audiovisual	MFCCs	GeMAPS	eGeMAPS	FAUs	Audiovisual
2 - class	SL v/s S	82.9	89.1	**93.0**	74.9	88.8	76.4	83.5	87.2	66.4	**88.2**
	AL v/s S	**99.5**	97.1	97.1	80.3	97.1	93.2	95.7	96.8	77.7	**96.9**
3 - class	UL/VL/SL	72.6	**74.9**	73.0	50.9	71.5	73.3	75.7	**77.0**	45.9	74.9
4 - class	UL/VL/SL/S	64.9	67.3	**72.5**	50.6	68.2	62.3	68.6	**73.4**	42.5	71.4

UL: Unvoiced Laughter, VL: Voiced Laughter, SL: Speech Laughter, AL: All Laughter, S: Speech.

standard-deviation were computed on the training partition and applied on both training and testing data (for each iteration of the LOSO loop). Performance of the classification model is measured as the unweighted average recall (UAR) of the classes. It is calculated by the sum of recall-values (class-wise accuracy) for all classes divided by the number of classes. This is the official scoring metric of the INTERSPEECH ComParE Challenge series [47]. Calculation of UAR is necessary to measure the correct performance, since in our case the class distribution is imbalanced; instances of the minority classes were upsampled to match the number of instances of the majority class found in the training partition. We evaluate the performance on the four feature sets: MFCCs, GeMAPS, eGeMAPS, and FAUs. Additionally, the best performing acoustic feature set is fused with the video feature set for audiovisual experiments (early-fusion).

4.3 Laughter Recognition

The aim in this set of experiments is to distinguish laughter from speech and distinguish between different kinds of laughter using the audio, visual and audiovisual features. For each feature set we perform four classification tasks: two binary classification tasks (Speech Laughter v/s Speech, and All laughter v/s Speech), one 3-class task for the discrimination of all kind of laughter categories, and a 4-class classification including all categories. We do not report results of binary classifications of voiced or unvoided laughter v/s speech as those laughter episodes were defined with semi-automatic rules and the obtained performance is obviously close to perfection.

In Table 6 we report the performance results obtained for both segmented and unsegmented set. For the segmented set, MFCCs performed best for the distinction between all types of laughter and speech instances, whereas the eGeMAPS acoustic feature set performed best for differentiating speech laughter from speech. Even though visual features performed much better than chance on the two binary classification tasks, they did not bring any additional improvement in the early fusion, since the acoustic features already provided very high recognition rates. For the unsegmented set, the eGeMAPS acoustic feature set performed the best for the 3-way and 4-way classification tasks, and the performance drops significantly when using only visual features. Interestingly, the combination of audio and visual data performs the best for the two binary classification tasks, with a slight improvement over the audio features, as also observed in previous studies [37, 38, 45]. From the results we can conclude that the different categories of laughter we defined can be well identified from speech segments using either audio or video feature set, and differentiating the categories of laughter performs better when using only the audio feature set since the information is mostly conveyed by the auditory channel.

4.4 Emotional Laughter Recognition

In this section we analyze the performance of automatic recognition of emotions in the different categories of laughter we studied. Binary emotion labels (negative/positive) were assigned to each class of laughter/speech based on the mean ratings of arousal and valence calculated using the delayed gold-standard. We first investigate how well the system can recognize emotions from individual categories of laughter and then perform the emotion classification tasks on all categories of laughter and all including speech.

Results obtained on the RECOLA dataset are reported in Table 7 and show that voiced laughter performed best for both arousal (77.1%) and valence (81.0%) when using audiovisual features; speech laughter performed equally well on arousal. Performance reported on speech laughter is generally slightly below the one reported for voiced laughter for valence. Results reported for unvoiced laughter show that such episodes convey much less emotion variability, especially for arousal, where only chance level is reported (50.0%). The performance obtained when using all instances of laughter is therefore lower because of the impact of unvoiced laughter. Interestingly, results obtained on speech utterances that do not include any laughter is slightly above the chance level for both arousal (52.8%) and valence (50.9%).

4.5 Context Recognition

Since laughter can be produced in a large variety of contextual situations, we wondered whether the context used in the data collections as reported in Section 2 could be automatically inferred from the audiovisual recordings. More specifically, we investigated two cases: *spontaneous laughter* v/s *acted laughter* (case 1), and *induced laughter* v/s *interactive laughter* (case 2).

We selected the MAHNOB [36] database for our first case since it provide both types of laughter produced by the same subjects, thus eliminating any bias due to inter-personal differences in the way laughter is produced. Further, we used the same dataset for training and testing to avoid the bias which can be due to different recording conditions, i. e., microphone, environment (room), external noise. As we can observe, the visual feature set (78.4%) performs better than the acoustic feature sets (75.6%) and the performance increases

Table 7: Results (%UAR) for 2-class (negative/positive) emotion classification for various categories of laughter using different feature acoustic and video feature sets; audiovisual corresponds to the early-fusion of the best acoustic feature set (eGeMAPS) and the video feature set; best results over the feature sets are highlighted in bold format.

Case	MFCCs		GeMAPS		eGeMAPS		FAUs		AudioVisual	
	Aro.	Val.	Aro.	Val.	Aro.	Val.	Aro.	Val.	Aro.	Val.
UL	50.0	59.4	50.0	58.2	50.0	**61.0**	50.0	60.7	**50.0**	54.2
VL	49.4	60.7	66.8	63.1	69.6	66.7	47.0	71.1	**77.1**	**81.0**
SL	75.4	63.6	74.1	71.1	**77.1**	69.0	75.7	**74.8**	**77.1**	74.5
AL	50.0	**61.1**	50.7	51.1	50.0	58.0	50.0	59.9	**50.2**	57.2
S	52.8	**50.9**	52.8	50.5	52.8	50.5	52.8	50.5	**52.8**	50.5
All	**51.8**	**50.6**	49.9	50.5	47.9	50.5	**51.8**	**50.6**	51.7	50.5

UL: Unvoiced Laughter, VL: Voiced Laughter, SL: Speech Laughter, AL: All Laughter, S: Speech.

significantly when both are combined (82.7%). The stereotype of laughter being associated with joy (Duchenne display) could explain the superiority of visual features over audio features for acted v/s spontaneous laughter distinction, since most of the acted laughter would involve the participants producing laughter with smile.

For the second case, we perform cross-corpora experiment since there is no database that provides both induced and interactive laughter instances. We fused the instances from the six datasets which include interactive data (RECOLA, SEMAINE, BELFAST) and induced data (MAHNOB, AVLC, BINED) and divided them into training and testing data (70:30 ratio). Results show that pure acoustic features (MFCCs) achieves a very high recognition rate, which might be helped by the acoustic variability present in the different corpora used (microphones, rooms), despite applying a z-score on the features for each dataset. However, the information extracted from the face, which is subject to less cross-corpora variabilities compared to speech, shows that performance is far above the chance level (79.5%), and even slightly better than in the first case (acted v/s spontaneous).

Results from these experiments thus show that there is a significant variability in the audiovisual expressions of laughter between acted, spontaneous, induced and interactive conditions.

Table 8: Results (%UAR) for 2 class classification task between Spontaneous and Acted laughter from MAHNOB and for laughter context recognition from speech (2 class) with six datasets using different feature sets : 3 audio, 1 video and the combination of best audio (MFCCs) and video feature set; best results over the feature sets are highlighted in bold format.

Case	MFCCs	GeMAPS	eGeMAPS	FAUs	Audiovisual
1	75.6	72.5	72.3	78.4	**82.7**
2	92.6	93.0	**93.9**	79.5	92.2

5 CONCLUSION

We have provided insights on the automatic analysis of emotional laughter by extensive evaluations carried out on the RECOLA database, which includes spontaneous interactions annotated in terms of time- and value-continuous emotional dimensions (arousal and valence). Annotations of laughter have been performed on this dataset, and will be made publicly available to the research community. We have then evaluated how the different annotated categories of laughter, such as unvoiced laughter, voiced laughter, speech laughter, and speech (non-laughter), can be automatically differentiated from audiovisual features, where very high recognition rates have been reported for various acoustic feature sets. Further, we have performed emotion recognition experiments on each of those categories. Results have shown that voiced laughter contains most of the emotion variabilities for both arousal and valence in classification tasks, i. e., passive vs. active, and negative vs. positive. Future work will investigate how variabilities in the language and culture might impact performance on the automatic recognition of laughter.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the University of Fribourg, Switzerland, through the commercial licenses of the RECOLA project and partial funding from the Council of Coaches project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant Agreement Number 769553. The authors would like to thank Dr. Catherine Pelachaud, ISIR-CNRS, Sorbonne Université for the support provided during the writing process of this research paper.

REFERENCES

[1] J. Bachorowski, M. Smoski, and M. Owren. 2001. The acoustic features of human laughter. *The Journal of the Acoustical Society of America* 110, 3 (2001), 1581–1597.

[2] J.-A. Bachorowski and M.J. Owren. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12, 3 (2001), 252–257.

[3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Proc. 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–6.

[4] S. Batliner, A.and Steidl, F. Eyben, and B. Schuller. 2011. On laughter and speech laugh, based on observations of child-robot interaction. *The phonetics of laughing* (2011).

[5] L.S. Berk, S.A. Tan, B.J. Napier, and W.C. Eby. 1989. Eustress of mirthful laughter modifies natural-killer cell activity. In *Proc. Clinical Research*, Vol. 37. 115A.

[6] P. Boersma and D. Weenink. 2005. Praat: doing phonetics by computer (version 4.3.01). In *Technical Report.* www.praat.org.

[7] D. Bone, C.-C. Lee, and S.S. Narayanan. 2014. Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features. *IEEE Transactions on Affective Computing* 5, 2 (April-June 2014), 201–213.

[8] H. Brugman and A. Russel. 2013. Annotating multimedia/multi-modal resources with ELAN. In *Proc. International Conference on Language Resources and Evaluation.* ELRA, 2065–2068.

[9] W.L. Chafe. 2007. *The importance of not being earnest: The feeling behind laughter and humor.* Vol. 3. John Benjamins Publishing.

[10] A.J. Chapman. 1976. Social aspects of humourous laughter. *Humour and laughter: Theory, research and applications* (1976), 155–185.

[11] S. Cosentino, S. Sessa, and A. Takanishi. 2016. Quantitative laughter detection, measurement, and classification – A Critical Survey. *IEEE Reviews in Biomedical engineering* 9 (2016), 148–162.

[12] C. Darwin, P. Ekman, and P. Prodger. 1998. *The expression of the emotions in man and animals.* Oxford University Press, USA.

[13] L. Devillers and L. Vidrascu. 2007. Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Proc. Interdisciplinary workshop on the phonetics of laughter.* 37.

[14] P. Ekman. 1997. What we have learned by measuring facial behavior. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)* (1997), 469–485.

[15] P. Ekman, R.J. Davidson, and W.V. Friesen. 1990. The Duchenne smile: Emotional expression and brain physiology: II. *Journal of personality and social psychology* 58, 2 (1990), 342.

[16] P. Ekman, W. V. Friesen, and J.C. Hager. 2002. *Facial action coding system.* Salt Lake City, UT: Research Nexus.

[17] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. 2015. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* (2015). in press.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia (MM).* ACM, 835–838. https://doi.org/10.1145/2502081.2502224

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* 9, Aug (2008), 1871–1874.

[20] P. Glenn. 2003. *Laughter in interaction.* Vol. 18. Cambridge University Press.

[21] J. Hall and W.H. Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human relations* 23, 4 (1970), 299–317.

[22] A. Hanjalic and L. Xu. 2001. User-oriented affective video content analysis. In *IEEE Workshop on Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001).* IEEE, 50–57.

[23] W. Hudenko, W. Stone, and J.-A. Bachorowski. 2009. Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder. *Journal of autism and developmental disorders* 39, 10 (2009), 1392–1400.

[24] K. Laskowski. 2009. Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.* IEEE, 4765–4768.

[25] K. Laskowski and S. Burger. 2007. Analysis of the occurrence of laughter in meetings.. In *Proc. INTERSPEECH.* 1258–1261.

[26] K. Laskowski and T. Schultz. 2008. Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. *Machine Learning for Multimodal Interaction* (2008), 149–160.

[27] F. Lingenfelser, J. Wagner, El. André, G. McKeown, and W. Curran. 2014. An event driven fusion approach for enjoyment recognition in real-time. In *Proc. of the 22nd ACM international conference on Multimedia.* ACM, 377–386.

[28] R.A. Martin. 2001. Humor, laughter, and physical health: Methodological issues and research findings. *Psychological bulletin* 127, 4 (2001), 504.

[29] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André. 2015. The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In *Proc. 2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015.* IEEE Computer Society, 166–172.

[30] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.

[31] C.C. Neuhoff and C. Schaefer. 2002. Effects of laughing, smiling, and howling on mood. *Psychological Reports* 91, 3_suppl (2002), 1079–1080.

[32] R. Niewiadomski, M. Mancini, T. Baur, G. Varni, H. Griffin, and M.S.H. Aung. 2013. MMLI: Multimodal multiperson corpus of laughter in interaction. In *Proc. International Workshop on Human Behavior Understanding.* Springer, 184–195.

[33] E. Nwokah, H.-C. Hsu, P. Davies, and A. Fogel. 1999. The integration of laughter and speech in vocal communication: A dynamic systems perspective. *Journal of Speech, Language, and Hearing Research* 42, 4 (1999), 880–894.

[34] F. Orozco, F. García, L. Arcos, and J. Gonzàlez. 2007. Spatio-temporal reasoning for reliable facial expression interpretation. In *Proc. International Conference on Computer Vision Systems (ICVS).* Bielefeld University.

[35] S. Petridis, M. Leveque, and M. Pantic. 2013. Audiovisual detection of laughter in human-machine interaction. In *Proc. 5th International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 129–134.

[36] S. Petridis, B. Martínez, and M. Pantic. 2013. The MAHNOB laughter database. *Image and Vision Computing* 31, 2 (2013), 186–202.

[37] S. Petridis and M. Pantic. 2008. Audiovisual laughter detection based on temporal features. In *Proc. of the 10th international conference on Multimodal interfaces.* ACM, 37–44.

[38] S. Petridis and M. Pantic. 2011. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE Transactions on Multimedia* 13, 2 (2011), 216–234.

[39] R.R. Provine. 2001. *Laughter: A scientific investigation.* Penguin.

[40] F. Ringeval, S. Amiriparian, F. Eyben, K.Scherer, and B. Schuller. 2014. Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion. In *Proc. of EmotiW, ICMI.* ACM, Istanbul, Turkey, 473–480.

[41] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. 2015. Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. *Pattern Recognition Letters* 66 (November 2015), 22–30.

[42] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge.* ACM, 3–8.

[43] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 2013. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proc. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).* IEEE.

[44] W. Ruch and P. Ekman. 2001. The expressive pattern of laughter. *Emotion, qualia, and consciousness* (2001), 426–443.

[45] S. Scherer, F. Schwenker, N. Campbell, and G. Palm. 2009. Multimodal laughter detection in natural discourses. In *Human Centered Robot Systems.* Springer, 111–120.

[46] M. Schröder. 2003. Experimental study of affect bursts. *Speech communication* 40, 1 (2003), 99–116.

[47] B. Schuller. 2012. The computational paralinguistics challenge [social sciences]. *IEEE Signal Processing Magazine* 29, 4 (2012), 97–101.

[48] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. 2012. The Belfast induced natural emotion database. *IEEE Transactions on Affective Computing* 3, 1 (2012), 32–41.

[49] R. Stibbard. 2000. Automated extraction of ToBI annotation data from the Reading/Leeds emotional speech corpus. In *Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.*

[50] M. T. Suarez, J. Cu, and M. Sta. 2012. Building a Multimodal Laughter Database for Emotion Recognition.. In *Proc. LREC.* 2347–2350.

[51] D.P. Szameitat, K. Alter, A.J. Szameitat, C.J. Darwin, D. Wildgruber, S. Dietrich, and A. Sterr. 2009. Differentiation of emotions in laughter at the behavioral level. *Emotion* 9, 3 (2009), 397.

[52] J. Trouvain. 2001. Phonetic Aspects of Speech-Laughs. In *Proc. of the Conference on Orality & Gestuality (ORAGE).* 634–639.

[53] J. Trouvain. 2003. Segmenting Phonetic Units in Laughter. In *Proc. of the 15th International Conference of Phonetic Sciences.* Barcelona Spain, 2793–2796.

[54] K. Truong and J. Trouvain. 2012. Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis. *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals* (2012), 20–24.

[55] J. Urbain. 2014. *Acoustic Laughter Processingn.* Ph.D. Dissertation. University of Mons.

[56] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. 2010. The AVLaughterCycle Database.. In *Proc. LREC.*

[57] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. 2016. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge.* ACM, 3–10.

[58] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. 2014. AVEC 2014 – The Three Dimensional Affect and Depression Challenge. In *Proc. of ACM MM.* Orlando (FL), USA.