Tracking authentic and in-the-wild emotions using speech

**Vedhas Pandit, Nicholas Cummins, Maximilian Schmitt, Simone Hantke,
Franz Graf, Lucas Paletta, Björn Schuller**

Tracking Authentic and In-the-wild Emotions Using Speech

Vedhas Pandit
ZD.B Chair of EIHW
University of Augsburg
Augsburg, Germany

Nicholas Cummins
ZD.B Chair of EIHW
University of Augsburg
Augsburg, Germany

Maximilian Schmitt
ZD.B Chair of EIHW
University of Augsburg
Augsburg, Germany

Simone Hantke
ZD.B Chair of EIHW
University of Augsburg
Augsburg, Germany

Franz Graf
Joanneum Research Forschungsgesellschaft mbH
Graz, Austria

Lucas Paletta
Joanneum Research Forschungsgesellschaft mbH
Graz, Austria

Björn Schuller
ZD.B Chair of EIHW
University of Augsburg
Augsburg, Germany
Group on Language, Audio, and Music (GLAM)
Imperial College
London, UK

Abstract—This first-of-its-kind study aims to track authentic affect representations in-the-wild. We use the 'Graz Real-life Affect in the Street and Supermarket (GRAS^2)' corpus featuring audiovisual recordings of random participants in non-laboratory conditions. The participants were initially unaware of being recorded. This paradigm enabled us to use a collection of a wide range of authentic, spontaneous and natural affective behaviours. Six raters annotated twenty-eight conversations averaging 2.5 minutes in duration, tracking the arousal and valence levels of the participants. We generate the gold standards through a novel robust Evaluator Weighted Estimator (EWE) formulation. We train Support Vector Regressors (SVR) and Recurrent Neural Networks (RNN) with the low-level-descriptors (LLDs) of the ComParE feature-set in different derived representations including bag-of-audio-words. Despite the challenging nature of this database, a fusion system achieved a highly promising concordance correlation coefficient (CCC) of .372 for arousal dimension, while RNNs achieved a top CCC of .223 in predicting valence, using a bag-of-features representation.

Index Terms—Affective Speech Analysis, Affective Computing, In-the-Wild, Authentic Emotions, Bag-of-Audio-Words, Gated Recurrent Units

I. INTRODUCTION

Contemporary research in human affect detection systems is often based on datasets collected in controlled settings, in which the participants are aware of being recorded, and of having their behaviours monitored [1]–[4]. This awareness has a detrimental effect on the authenticity and/or naturalness of the data collected. This effect is often referred to as the 'one-way mirror dilemma', or the 'observer's paradox' in the sociology literature [1], [2]. Therefore, the collection of real-world, natural and spontaneous data has long been regarded as a key step in improving human affect and sentiment detection

systems [3], [4]. To this end, we present – for the very first time – audio-based investigations on the temporal tracking of *authentic* or 'observers paradox'-free affective states collected in-the-wild.

We use the *'Graz Real-life Affect in the Street and Supermarket (GRAS^2)'* corpus [5], featuring truly spontaneous examples of human behaviours. The recordings are highly realistic; they were made at a public place – a busy shopping mall in Graz, Austria. They feature many real-life, 'non-laboratory' effects such as the varying background noises, accompanying people speaking, also the spontaneous, impromptu behaviours of the participants, all contributing to 'in-the-wild' nature of the data collected. Six raters annotated the arousal and valence levels of the subjects on-screen in these recordings. We compute the gold standards as the weighted sums of the individual annotations, where the weights represent the extent to which an annotation is in agreement with other annotations.

We aim to test effectiveness of conventional and deep learning approaches in tracking the authentic, in-the-wild affects, overcoming the new challenges posed by this unique data collection strategy. While affect detection is often performed using multimodal paradigms, the unconventional nature of the exchanges dictates us to focus on speech. This is because, majority of the frames do not fully feature the participant's face, as the assistants collecting the data often did not make an eye contact. Speech-based emotion detection is a widely used, robust alternative to multimodal systems [6]–[8].

In this regard, we use low-level descriptors from the INTERSPEECH *Computational Paralinguistics ChallengE* (COMPARE) feature-set [9], along with functionals such as mean, standard deviation, delta differentiation, and bag-of-

audio words (BoAW) which have proven to be particularly useful in the affect recognition tasks [10], [11]. We use *Support Vector Regressors* (SVRs) – conventionally the baseline method for regression tasks in COMPARE [12]–[14]; and *Recurrent Neural Networks* (RNN) – particularly useful for temporal sequence modelling [15], [16]. We fuse together predictions from the two models by computing a weighted sum, where the weights are proportional to concordance correlation coefficient (CCC) [17] achieved during training.

We train the models to analyse the real-world signals. Because the 'noise'-profile is continuously varying with the impact sounds, music and people in the background, there is no reliable way to compute signal-to-noise ratio (SNR) values. The issue is exacerbated by the recurring episodes of absence of speech, e. g., as the participant reads through the documents. Likewise, there is no clean speech or ground truth to make the source-separation based SNR calculations. Most importantly, as the noises present in the recordings are real-world, distorting and corrupting the signal with artificial noises for SNR values goes against the very purpose and spirit of this unique study.

The rest of the paper is organised as follows. Section II outlines the unique data collection paradigm of the GRAS^2 corpus. In Section III and Section IV, we present the experiments and results for the chosen emotion prediction regression task. Finally, in Section V, we conclude with potential future work directions.

II. THE DATABASE

The GRAS^2 corpus has previously been used in a study that proposed a methodology to detect the 'primary' speech segments (i. e., excluding speech segments from people in the background), using the correlations between acoustic cues and the visual attention [5]. Four students collected this data at the supermarket, working as research assistants. The *assistants* were equipped with audio recorders and eye tracking glasses featuring a frontal camera, along with other sensors recording physiological data [5]. With this setup, impromptu conversations with unsuspecting female shoppers were collected in both the video and audio modalities.

The assistants pretended to be searching for a particular store or a product, and sought help from their dialogue partners – referred to as *participants* from here onwards. The peculiar choice of products, coupled with continued requests for help, as well as the revelation of them being recorded in an experiment, were all intended to elicit a range of emotional behaviours (e. g., disgust, confusion, surprise, laughter). The recordings were preserved and studied only upon obtaining the consent from the participants. Audio segments featuring speech from anyone accompanying the participant were later muted, consistent to data collection and privacy policy. Table I presents an overview of the dataset.

Whilst the interactions were unscripted, commonalities have been observed in these exchanges. The exchanges typically progressed as follows (Figure 1):

TABLE I: GRAS^2 statistics for the 28 conversations in terms of the durations and number of utterances by the assistant (Ass.) and the recorded participant (Part.).

	Duration	Num. of utterances			Gold Standard	
		Total	Ass.	Part.	Arousal	Valence
Minimum	71 sec	27	16	9	-.200	-.238
Maximum	309 sec	111	54	67	.570	.564
Average	142.9 sec	52.1	27.9	24.2	.144	.105
Std.Dev.	57.3 sec	21.3	9.7	12.8	.103	.115
Total	66 mins 42 sec	1458	780	678	–	–

(a) Participant interrupted

(b) Participant responds

(c) The revelation

(d) Documentation presented
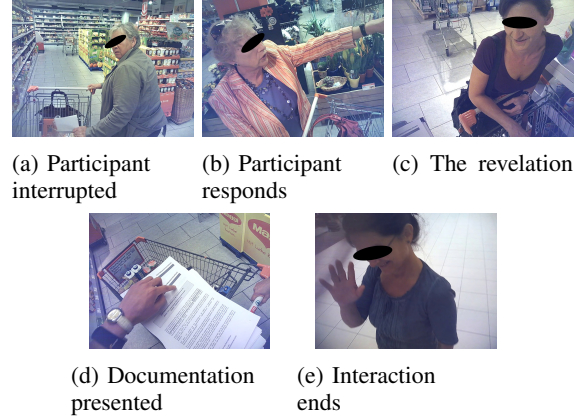
(e) Interaction ends

Fig. 1: All exchanges in the $GRAS^2$ dataset are unscripted, however the depicted five events are common throughout

1) *Participant interruption:* First, a participant gets interrupted on the way. Depending on the recording location, is asked about some products, or a shop (Figure 1a).
2) *Participant responds:* The participant responds, the assistant often asks follow up questions (Figure 1b).
3) *The revelation:* The assistant makes the participant aware of being recorded. This phase may feature even more of the impromptu conversations, unpredictable interruptions coming from the participant (Figure 1c).
4) *Documentation presented:* The assistant hands the paperwork to the participant which includes a consent form, and an optional BFI-10 survey [18] with questions that help assess their personality. The duration of this phase and amount of conversions varies a lot between the participants (Figure 1d).
5) *Interaction ends:* The conversation ends, often after some exchange of greetings (Figure 1e).

III. EXPERIMENTS

A. Gold Standard Generation

We collect the annotations using the crowdsourcing-based gamified platform iHEARu-PLAY [20]. Six annotators with different ethnic backgrounds and German language proficiency watched the GRAS^2 audiovisual recordings, and separately annotated the perceived arousal and valence levels of the interacting participants on-screen by dragging the slider between -1 and 1 with a step-resolution of 0.1 with a mouse. We use a novel *Evaluator Weighted Estimator* (EWE) method to
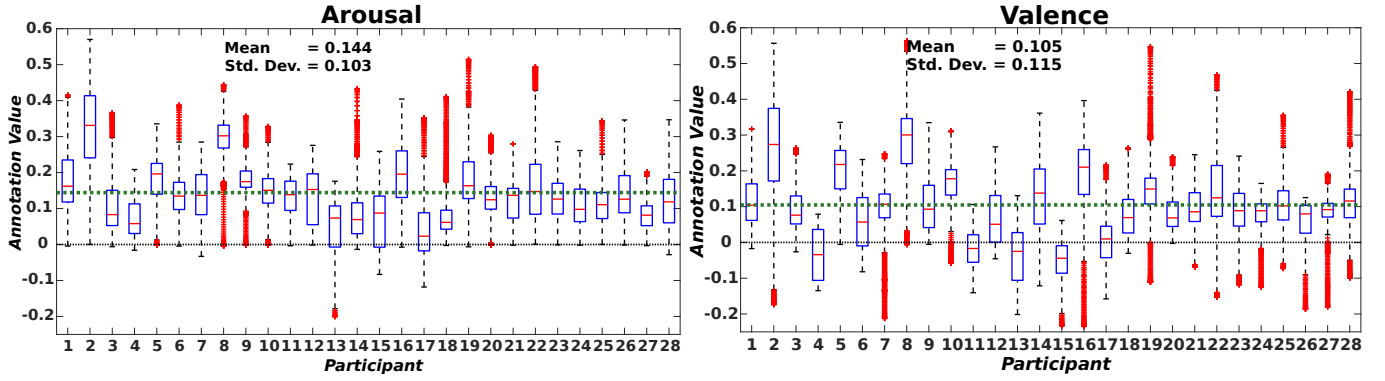
Fig. 2: Boxplots of the gold standards. Overall mean value shown in green, 0-level in black. Presence of large number of outliers (in red) indicates challenging nature of the data. The outliers are determined by the MATLAB's default settings for its inbuilt *boxplot()* function. The arousal annotations are mostly positive. While valence gold standard does feature some negative values, annotations are mostly positive in both the emotion dimensions, evident from the y-axis range and the quartiles in blue.
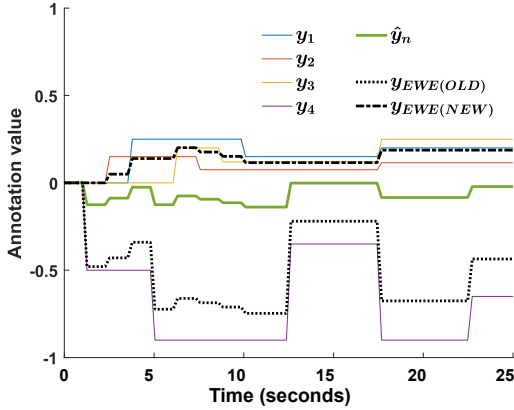

Fig. 3: A synthetic set of annotations used to highlight the effect of outliers on the calculation of a gold standard rating. Note that the gold standard calculated using $y_{EWE(OLD)}$, [19], is heavily influenced by annotation y_4 whilst the gold standard calculated using the methodology proposed in this paper $y_{EWE(NEW)}$ is closer aligned to the set of annotations in agreement with each other, namely y_1, y_2 and y_3

generate the *gold standard*s, one per participant per emotion dimension. The *EWE* metric takes into account confidence over the individual annotators, and assigns the weight r_k for every annotation y_k. The gold standard, y_{EWE} is defined by the equation:

$$y_{EWE_n} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k y_{n,k}, \quad (1)$$

where $y_{n,k}$ is the annotation by the annotator k ($k \in \mathbb{N}, 1 \leq k \leq K$) at instant n ($n \in \mathbb{N}, 1 \leq n \leq N$) contributing to the annotation sequence y_k, and r_k is the corresponding annotator-dependent weight. The lower bound for r_k is set to 0.

In [19], the weight r_k is defined to be normalised cross-correlation between y_k and the averaged annotation sequence $\bar{y}_n$, . The computation of r_k, thus, depends largely on the $\bar{y}_n$,

assumed to be a good representative of the sequences y_k. The $\bar{y}_n$ sequence however, is easily influenced by the large absolute values in y_k, and not necessarily by the extent to which the sequences in y_k are correlated with one other. Thus, contrary to the expectations, $\bar{y}_n$ can potentially become very similar to an outlier sequence. With such a formulation, $\bar{y}_n$ can in theory become very similar to the outlier sequence, contrary to becoming the representative sequence of the correlated y_k. A synthetic example of this effect is given in (Figure 3).

We therefore redefine the weight r_k such that it gets strongly influenced by the total number of annotations y_k is in agreement with, and also by the extent to which they agree, by simply averaging the pair-wise correlations:

$$r'_{k_i,k_j} = \frac{\sum_{n=1}^{N} (y_{n,k_i} - \mu_{k_i})(y_{n,k_j} - \mu_{k_j})}{\sqrt{\sum_{n=1}^{N} (y_{n,k_i} - \mu_{k_i})^2}\sqrt{\sum_{n=1}^{N} (y_{n,k_j} - \mu_{k_j})^2}},$$

$$\text{where: } \mu_k = \frac{1}{N}\sum_{n'=1}^{N} y_{n',k}, \quad (2)$$

$$r_{k_i} = \begin{cases} \frac{1}{K}\sum_{k_j=1}^{K} r'_{k_i,k_j} & \text{if } \sum_{k_j=1}^{K} r'_{k_i,k_j} > 0 \\ 0 & \text{if } \sum_{k_j=1}^{K} r'_{k_i,k_j} \leq 0 \end{cases}. \quad (3)$$

This modified weight computation is similar to that presented in [21], except that we include the autocorrelation values which diminishes the effect of the outlier annotation that is extremely negatively correlated [14]. Statistics and distributions of the per-participant gold standard scores are as shown in Figure 2.

B. Feature Representations

People voluntarily or involuntarily communicate their emotions through facial expressions and verbal/non-verbal vocalisations [4]. Due to the in-the-wild nature of the dataset, more than two thirds of the video frames on average do not feature the participant's face, as the assistant often swerves away to look at the documents, floor, background objects. Visual
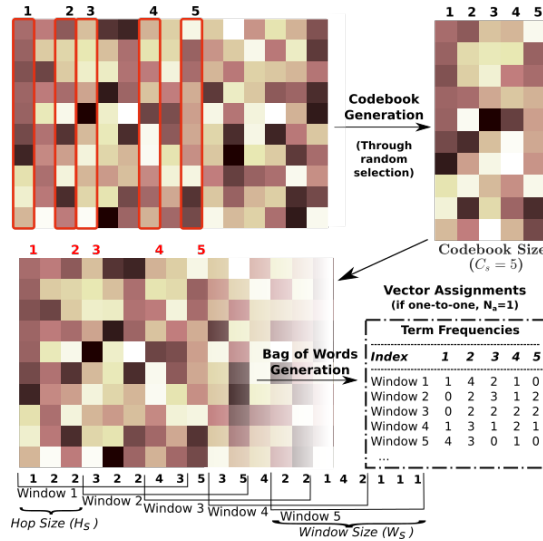
Fig. 4: BoAW feature generation exemplified [24], [25]. In the example above, $C_s = 5$ features are randomly selected to create a codebook. Each feature vector is then vector-quantised to those in the codebook. The sequential dynamic variation in the distribution of quantised feature vectors is captured by taking the histogram of the assignments in each window.

features therefore could not be used readily. Accordingly, we use audio features to predict arousal and valence dimensions of the emotions expressed, testing relevance of both the short-term features, i. e., the Low Level Descriptors (LLDs) and long-term features (functionals computed over LLDs).

Our first audio feature-set is made up of 130 LLDs: the 65 basic LLD features and their first order derivatives (deltas), contained in the COMPARE feature-set [9]. This feature-set includes prosodic and spectral features (e. g., F_0, Sum of RASTA-filtered auditory spectrum, MFCCs), as well as those related to voice quality (e. g., log HNR, probability of voicing, jitter), and is known to be suitable for affect prediction [21], [22]. The LLDs were extracted using the openSMILE toolkit [23] with a window size of 25 ms and a hop size of 10 ms.

Due in part to memory constraints, the full COMPARE feature-set (a 6373-dimensional feature vector) for every time step was computationally infeasible to train on. However, many of the functionals used in COMPARE make little contribution for emotion prediction for small window sizes (e. g., minPos, maxPos) [9], [22]. We therefore adapt a minimalist approach by computing only mean and standard deviations of the 65 LLDs. These functionals are computed over a window size of 1 second, with a hop of 10 ms. We match the sampling period of the annotations (100 ms) by applying a moving average filter on 10 feature vectors at a time, with a hop size of 100 ms to both the functional features and the LLD features.

The *Bag-of-Audio Words* (BoAW) feature representation has achieved state-of-art performance for emotion detection on the RECOLA dataset outperforming techniques such as End-to-End learning [24]. BoAW involves generation of a sparse fixed length histogram representation of an audio instance, similar to

the popular *Bag-of-Words* paradigm from the natural language processing field. The histogram represents the frequency of each identified audio word in a given audio instance (Figure 4). Due in part to it's inherent sparsity and the quantisation step, the BoAW representation is inherently less sensitive to the individual feature vectors computed over a small window size, e. g., 25 ms or 100 ms. This feature transformation instead captures the varying temporal trends in the distribution of quantizations of these input feature vectors in a much larger time-scale, e. g., 5 to 10 seconds. It is therefore considered to be a more robust representation than the LLDs or the functionals [24], [25].

We compute the BoAW features using the open-source openXBOW toolkit [25]. Initial experiments on the hyperparameters (results not given) revealed the most suitable setup for our needs to be a codebook of randomly chosen 100 features ($C_S = 100$). 1 assignment of every feature vector ($N_a = 1$) was computed every 100 ms ($H_s = 100$ ms) to match the annotation sampling period, for an 8 second long moving window ($W_s = 8$ seconds).

It is well known that compensating for *annotation delay* improves system performance [13], [24], [26], [27]. Initial experimentation with values ranging from 2.5 seconds to 4.5 seconds revealed, that lag compensation of 3 seconds was suitable for the GRAS^2 corpus; consistent with the results indicating that the annotation lag varies between 1 to 6 seconds [26], [27].

C. Time-continuous Regression Models

Due to limited sample size (28 videos), we present the results for the leave-one-session-out (LOSO) cross-validation using the Support Vector Regressor (SVR) and the GRU-based Recurrent Neural Network (GRU-RNN). The support vector regressions were implemented using the sklearn package [28], while *gated recurrent unit* (GRU)-based recurrent neural networks were trained using the tensorflow package [29]. The results are reported in terms of the mean and standard deviation of the *Concordance Correlation Coefficient* (CCC) metric [17], calculated from the results for each fold.

The SVR model involves several hyperparameters that directly impact model's generalisability, speed, and convergence. In our experiments (cf. Section IV), we test a range of *complexity parameters* (C) from $[10^{-3}, 10^{-2}, ... , 10^3]$ with either a linear, polynomial (degree of 3) or radial basis function kernel and *epsilon* (ε) values chosen from $[10^{-3}, 10^{-2}, 10^{-1}]$.

The SVR paradigm does not capture sequential context and long term dependencies in the data. Among a wide array of neural network architectures, RNNs are arguably the best suited to do this task; we therefore assess the suitability of *Gated Recurrent Unit* (GRU)-based RNN framework. We feed the sampling-period matched (100 ms) input features to a two-layered GRU-RNN, the most recent output of which is then used to predict the emotion-regression value using a three-layer deep fully connected neural network. We experimented with different number of nodes for each layer (e. g., 30, 50, 100 nodes for the GRU layers), different activation function

TABLE II: The best performing configurations for SVR and GRU-RNN models, for different feat(ure) rep(resentations). ED: Emotion Dimension, A: Arousal, V: Valence, K: Kernel, L: Linear kernel, P: Polynomial kernel (degree 3), C: Complexity, FFL: Feedforward layer.

ED	Feat. Rep.	SVR K	C	ε	GRU-RNN
A	LLDs	L	10^{-2}	10^{-3}	50×50 GRU (tanh)
	Functionals	L	10^{-2}	10^{-3}	$\times 15$ FFL (linear)
	Bag-of-LLDs	P	10^{+3}	10^{-2}	$\times 4$ FFL (linear)
	Bag-of-Func.	P	10^{+3}	10^{-2}	$\times 1$ FFL (tanh)
V	LLDs	L	10^{-2}	10^{-1}	35×35 GRU (tanh)
	Functionals	L	10^{-2}	10^{-1}	$\times 15$ FFL (linear)
	Bag-of-LLDs	P	10^{+3}	10^{-2}	$\times 4$ FFL (linear)
	Bag-of-Func.	P	10^{+3}	10^{-2}	$\times 1$ FFL (tanh)

combinations (e. g., *tanh* and *linear*), and also in terms of reducing the number of layers. Another hyper-parameter for the neural network is the number of training iterations it goes through, which directly dictates its performance during both the training and the testing phases. We also compute weighted sum of the two predictions, where weights are proportional to CCC achieved during training. To minimise the risk of overfitting, we terminate the training for every fold whenever the CCC for the concatenated training predictions of more than 0.7 is reached (empirically chosen vale through grid-search experiments on aforementioned hyper-parameters). For the folds where a CCC of 0.7 is not reached even after 1000 epochs of training, we choose the trained model with the highest recorded CCC value on the training data.

Table II lists the best performing hyper-parameters corresponding to the results presented in Section IV.

IV. RESULTS

While the mean CCC for the GRU-RNN system was observed to be in the range from 0.136 to 0.223 for valence prediction, and from 0.143 to 0.370 for arousal prediction (Table III), it goes well beyond 0.7 for some of the individual folds on the test data. This is especially true for the test clips where the participants spoke more, and where the audio content was more expressive in terms of the speech, and the emotional responses such as the laughter.

The weaker performance of the LLDs could be due in part to the challenging nature of this dataset (Figure 2). The recordings contain many non-speech episodes, many and varying background sounds; featuring also the background speech, music, impact sounds. The short term features are likely to capture irrelevant audio events, whereas features computed over longer frames are likely to smooth out the irrelevant information. This conjecture has been verified in part by the comparatively stronger performance of the functional features which are computed over a longer time frame performed better than the short-term LLDs. Similarly, as discussed in Section III-B, due to vector quantisation, the bag-of-features approach is even less sensitive to the noise in individual feature vectors.

TABLE III: Leave-one-participant-out cross-validation performance results, in terms of the Concordance Correlation Coefficient (CCC). The results are given in terms of the different regression paradigms; Support Vector Regression (SVR) and Gated Recurrent Unit based Recurrent Neural Networks (GRU- RNNs); the arousal and valence emotion dimensions; and the different Feat(ure) Rep(resentations).

Model	Dimension	Feat. Rep	Mean CCC	Std. Dev. CCC
SVR	Arousal	LLDs	.122	.106
		Functionals	.232	.164
		Bag-of-LLDs	**.327**	.208
		Bag-of-Func.	.178	.126
	Valence	LLDs	.055	.094
		Functionals	.123	.137
		Bag-of-LLDs	**.162**	.184
		Bag-of-Func.	.067	.819
GRU-RNN	Arousal	LLDs	.189	.227
		Functionals	.143	.284
		Bag-of-LLDs	**.370**	.237
		Bag-of-Func.	.328	.203
	Valence	LLDs	.136	.175
		Functionals	.213	.235
		Bag-of-LLDs	.191	.216
		Bag-of-Func.	**.223**	.201
Weighted Sum	Arousal	LLDs	.196	.225
		Functionals	.185	.244
		Bag-of-LLDs	**.372**	.231
		Bag-of-Func.	.311	.190
	Valence	LLDs	.138	.165
		Functionals	**.219**	.224
		Bag-of-LLDs	.216	.212
		Bag-of-Func.	.203	.181

RNNs fair a lot better than SVRs, as they capture long term temporal dependence of the features – an important trait for emotion prediction [8], [16]. The estimation of valence was, unsurprisingly, more difficult to model than the arousal using audio only. This is consistent with findings reported in previous studies [24], [30].

For arousal prediction, the fusion of the different classification system generally improved prediction performance of the unfused systems (cf. Table III). The fused Bag-of-Functional system gained our strongest arousal CCC of 0.372. The advantages of fusion for valence prediction are less clear (cf. Table III). In general, the valence fusion systems were unable to outperform the (unfused) GRU-RNN systems. We speculate that this is due to the weaker unfused SVM predictions having a negative effect on the fusion.

When using only the acoustic features, one of the biggest challenges is the prevalence of long non-speech segments in GRAS^2 dataset, when the annotators are likely to have marked the perceived arousal and valence levels relying mostly on the facial expressions or gestures. Moreover, being a first of its kind study on authentic emotion representations, the mean CCC's observed are not directly comparable against emotion tracking studies done on *controlled* datasets.

V. CONCLUSIONS AND FUTURE WORK

In this work, we observed the effectiveness of RNNs over SVR models for speech-based emotion prediction in both the

arousal and valence dimensions on a unique and challenging dataset. In terms of the features, the bag-of-features approach works well for the most part when compared to individual LLDs or the functionals. This is most likely due to quantisation procedure making the resulting feature representation less sensitive to noise. We also observed that for arousal predictions fusion brings together best of the both models, resulting in an improved prediction performance.

Because the dataset contains point-of-view (PoV) audio-visual recordings, it is difficult to track the participants' face in its entirety for majority of the frames. As the facial expression information is crucial, incorporating the video modality remains a big challenge particular to this dataset. As part of the future work, we intend to alleviate the problem by implementing multimodal result fusion assigning smalls weights to the video frame-based results when no face gets detected. We intend to make use of physiological data streams, such as assistant's electrodermal activity, gaze to investigate the correlations in predicting the participant's affect states.

REFERENCES

[1] S. Speer and I. Hutchby, "From Ethics to Analytics: Aspects of Participants' Orientations to the Presence and Relevance of Recording Devices," *Sociology*, vol. 37, no. 2, pp. 315–337, 2003.

[2] I. Hutchby, M. O'Reilly and N. Parker, "Ethics in praxis: Negotiating the presence and functions of a video camera in family therapy," *Discourse Studies*, vol. 14, no. 6, pp. 675–690, 2012.

[3] R. W. Picard, E. Vyzas and J. Healey, "Toward Machine Emotional Intelligence:Analysis of Affective Physiological State," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[4] M. Pantic, N. Sebe, J. F. Cohn and T. Huang, "Affective Multimodal Human-computer Interaction," in *Proc. 13th ACM MM*, Singapore, Nov. 2005, pp. 669–676, ACM.

[5] F. Eyben, F. Weninger, L. Paletta and B. Schuller, "The acoustics of eye contact – Detecting visual attention from conversational audio cues," in *Proc. 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction (GAZEIN) at 15th ICMI*, Sydney, Australia, Dec. 2013, pp. 7–12, ACM.

[6] M. El Ayadi, M. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut et al, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th ICMI*, State College,PA, 2004, pp. 205–211, ACM.

[8] F. Weninger, F. Ringeval, E. Marchi and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio," in *Proc. 25th IJCAI*, New York City, NY, Jul. 2016, pp. 2196–2202.

[9] B. Schuller, S. Steidl, A. Batliner, A.Vinciarelli et al, "The INTER-SPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. 14th INTERSPEECH* , Lyon, France, Aug. 2013, pp. 148–152, ISCA.

[10] C. Anagnostopoulos, T. Iliou and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intell. Review*, vol. 43, no. 2, pp. 155–177, 2015.

[11] J. Deng, N. Cummins, J. Han, X. Xu et al, "The University of Passau Open Emotion Recognition System for the Multimodal Emotion Challenge," in *Proc. 7th CCPR*, Chengdu, P. R. China, Nov. 2016, pp. 652–666, Springer.

[12] F. Ringeval, B. Schuller, M. Valstar, R. Cowie et al, "AVEC 2015: The 5th International Audio/Visual Emotion Challenge and Workshop," in *Proc. 5th Int. Workshop on Audio/Visual Emotion Challenge (AVEC'15) at 23rd ACM MM*, Brisbane, Australia, Oct. 2015, pp. 1335–1336, ACM.

[13] M. Valstar, J. Gratch, B. Schuller, F. Ringeval et al, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proc. 6th Int. Workshop on Audio/Visual Emotion Challenge (AVEC'16) at 24th ACM MM*, Amsterdam, The Netherlands, Oct. 2016, pp. 3–10, ACM.

[14] F. Ringeval, B. Schuller, M. Valstar, S. Mozgai et al, "AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proc. 7th Int. Workshop on Audio/Visual Emotion Challenge (AVEC'17) at 25th ACM MM*, Mountain View, CA, Oct. 2017, pp. 3–9, ACM.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau et al, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP* , Doha, Qatar, Oct. 2014, pp. 1724–1734, ACL.

[16] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller et al, "LSTM-Modeling of Continuous Emotions in an Audiovisual Affect Recognition Framework," *Image and Vision Computing, Special Issue on Affect Anal. in Continuous Input*, vol. 31, no. 2, pp. 153–163, Feb. 2013.

[17] I Lawrence and Kuei Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989.

[18] Beatrice Rammstedt and Oliver P John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[19] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automat. Speech Recognition and Understanding (ASRU)*, Nov. 2005, pp. 381–385.

[20] S. Hantke, F. Eyben, T. Appel and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. 1st Int. Workshop on Automat. Sentiment Anal. in the Wild (WASA) at 6th ACII*, Xi'an, P.R. China, Sep. 2015, pp. 891–897, IEEE.

[21] F. Ringeval, F. Eyben, Eleni Kroupi, Anil Yuce et al, "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, Nov. 2015.

[22] F. Eyben, Klaus Scherer, B. Schuller, Johan Sundberg et al, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[23] F. Eyben, Felix Weninger, F. Groß and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. 21st ACM MM 2013*, Barcelona, Spain, Oct. 2013, pp. 835–838, ACM.

[24] M. Schmitt, F. Ringeval and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. 17th INTERSPEECH* , San Francisco, CA, Sep. 2016, pp. 495–499, ISCA.

[25] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *J. Mach. Learn. Res.*, vol. 18, 2017.

[26] Z. Huang, T. Dang, N. Cummins, B. Stasak et al, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. 5th Int. Workshop on Audio/Visual Emotion Challenge (AVEC'15) at 23rd ACM MM*, Brisbane, Australia, Oct. 2015, pp. 41–48.

[27] 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, Apr. 2015.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel et al, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo et al, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[30] F. Ringeval, Andreas Sonderegger, Juergen Sauer and Denis Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative andAffective Interactions," in *10th IEEE Int. Conf. and Workshops on Automat. Face and Gesture Recognition (FG'13)*, Shanghai, P. R. China, Apr., pp. 1–8.