How Good Is Your Model 'Really'? On 'Wildness' of the In-the-wild Speech-based Affect Recognisers

Vedhas Pandit[1][0000−0002−1983−8140], Maximilian Schmitt[1], Nicholas Cummins[1], Franz Graf[2], Lucas Paletta[2], and Björn Schuller[1,3]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Joanneum Research Forschungsgesellschaft mbH, Austria
[3] Group on Language, Audio, and Music (GLAM), Imperial College London, UK

Abstract. We evaluate, for the first time, the generalisability of in-the-wild speech-based affect tracking models using the database used in the 'Affect Recognition' sub-challenge of the Audio/Visual Emotion Challenge and Workshop (AVEC 2017) – namely the 'Automatic Sentiment Analysis in the Wild (SEWA)' and the 'Graz Real-life Affect in the Street and Supermarket (GRAS^2)' corpus. The *GRAS^2* corpus is the only corpus to date featuring audiovisual recordings and time-continuous affect labels of the random participants recorded surreptitiously in a public place. The *SEWA* database was also collected in an in-the-wild paradigm in that it also features spontaneous affect behaviours, and real-life acoustic disruptions due to connectivity and hardware problems. The *SEWA* participants, however, were well aware of being recorded throughout, and thus the data potentially suffers from the 'observer's paradox' . In this paper, we evaluate how a model trained on a typical data suffering from the observer's paradox (SEWA) fairs on a real-life data that is relatively free from such psychological effect (GRAS^2), and vice versa. Because of the drastically different recording conditions and the recording equipments, the feature spaces for the two databases differ extremely. The in-the-wild nature of the real-life databases, and the extreme disparity between the feature spaces are the key challenges tackled in this paper, a problem of a high practical relevance. We extract bag of audio words features using, for the very first time, a randomised database-independent codebook. True to our hypothesis, the Support Vector Regression model trained on GRAS^2 had better generalisability, as this model could reasonably predict the SEWA arousal labels.

Keywords: Affective Speech Analysis · Transfer Learning · Observer's Paradox · One-way Mirror Dilemma · Authentic Emotions · In-the-Wild

1 Introduction

Human speech is a complex signal, featuring a plethora of information beyond the spoken words. In addition to the linguistic content, a speech signal tells the

listener a lot about the speaker – such as their age, gender, native language, motivations and emotions. It is important for a human-machine interaction (HCI) system to recognise these contexts correctly, to be able to respond in accordance. Today, we are continuously surrounded by human-machine interfaces. A virtual assistant in a handheld device has no longer remained a science fiction, but is simply an everyday reality. There is, therefore, a growing interest in the field of affective computing, to make the machines 'understand' human speech in its entirety, i. e., including the featured emotions and contexts.

Broadly speaking, there are three types of databases used in affect research. Early research utilised acted speech data, which typically featured highly exaggerated affect behaviours, far from the natural ones (e. g., EmoDB [1, 2]). In another data collection strategy, the participants are made to converse in a laboratory environment. While the behaviours collected are mostly natural and spontaneous, the collected data is typically clean and unaffected by the real-life effects such as noise (e. g., RECOLA [3]). The third, 'in-the-wild' databases refer to the data collected in non-laboratory, the everyday, unpredictable noisy environments. However, the so-called 'in-the-wild' databases mostly feature the recordings collected in identical real-life settings, with very similar acoustic disruptions. This has direct implications on the trained models, limiting their generalisability. Also, most of these databases suffer from the phenomenon called 'observer's paradox' or 'one-way mirror dilemma' – where the participants are typically well aware of being recorded right from the beginning of the recordings – which affects featured affect behaviours [4]. In this contribution, we test, for the first time, the hypothesis that the models trained on a closer-to-real-life database is likely to generalise better [5].

While there have been transfer learning studies on affect [6–9], there is hardly any research on generalisability of time-continuous affect recognising models for the real-life or in-the-wild datasets. To this end, we first introduce the two databases used in this study in Section 2. We describe our experiments in detail in Section 3. After this, we present our findings in Section 4 before we conclude the paper in Section 5.

2 Databases

To test which of the two affect recognising models generalises better – i. e., whether the one trained on a 'more' in-the-wild database or the one using database collected under relatively restrained or 'laboratory'-like settings – we use two prominent benchmark databases, namely the 'Automatic Sentiment Analysis in the Wild' (SEWA) corpus used in the AVEC 2017 challenge and the 'Graz Real-life Affect in the Street and Supermarket' (GRAS^2) corpus.

The **SEWA** database features video chat recordings of the participants discussing the commercials they just watched. The recordings were collected using the standard webcams and computers from the participants' homes or offices. The data collection took place over the internet using a video-chat interface specifically designed for this task. The recordings feature spontaneous affect be-

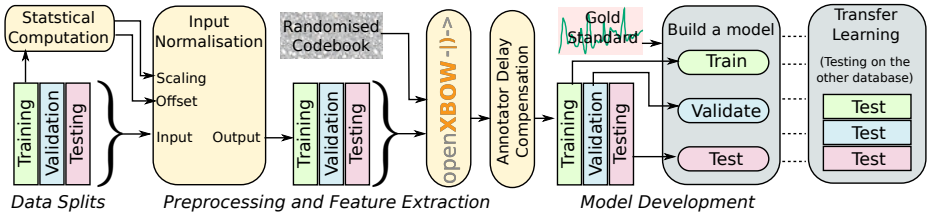

Fig. 1. Entire experimental design pipeline

haviours, real-life noises and delays due to connectivity and hardware problems. The participants dominate the conversations more or less equally in most cases.

The **GRAS^2** database features audiovisual recordings of the conversations with the unsuspecting participants from a first-person point of view in a busy shopping mall. The participants were made aware of being recorded only half way through the conversations, and were requested to sign a consent form agreeing to release the recordings for research purposes. The database, thus, features spontaneous and 'more' authentic affective behaviours, as they are relatively more observer's paradox-free. Because the conversations were totally spontaneous, the durations of the conversations vary widely (standard deviation = 56.3 seconds). Also the extent to which the participants dominate the conversations, i. e., relative durations of the participant's speech and the speech by the student research assistant collecting the data, varies widely. Unfortunately, the student research assistants dominate many of the conversations. The sections of the recordings where the participants read the documents before signing the consent form hardly feature participant's speech. The recordings also contain dynamically varying noise, including the impact sounds, bustle, background music, and background speech. There are only 28 conversations available. All these factors combine to make the this database a lot more 'in-the-wild' and the affect tracking task lot more challenging. The corpus was used previously in a research study establishing correlation between an eye-contact and the speech [10], and another study on time-continuous authentic affect recognition in-the-wild [11].

3 Experimental Design

3.1 Data Splits

We split both the SEWA and GRAS^2 corpus into training, validation and test sets in a roughly similar 2:1:1 ratio, in terms of both the number of files in a split and the cumulative duration of the audio clips. We use the same splits used in the AVEC 2017 challenge [12] when running our experiments on the SEWA database. The splits are made such that a participant-independent model can be trained, i. e., no participant is present in more than one split. The splits on GRAS^2 are made such that each split features a different student assistant likewise, i. e., no student assistant is present in more than one split. The statistics for the three splits are presented in Table 1.

Table 1. Duration statistics for the SEWA and GRAS^2 data splits

		SEWA			GRAS^2		
		Train	**Validation**	**Test**	**Train**	**Validation**	**Test**
Duration (seconds)	**Total**	5608.02	2272.30	2807.42	2018.75	1000.45	998.02
	Max	175.64	175.45	175.81	218.77	290.94	309.40
	Min	46.68	97.43	174.9	71.77	100.31	86.40
	Mean	164.94	162.31	175.46	126.17	166.74	166.34
	Std. Dev.	31.24	26.71	0.24	34.90	63.67	74.93
Number of Subjects		34	14	16	16	6	6

3.2 Feature Engineering

We need the features from the two databases such that they are compatible with one another, the two ideally share a common feature space. Because we are interested in predicting time-continuous signals of emotion dimensions, the features should also ideally capture the temporal dynamics of the varying low-level descriptor (LLD) space. The features should ideally be robust to noise.

We generate the bags of audio words (BoAW) features using our own openXBOW toolkit [13] by vector quantising the 'enhanced Geneva Minimalistic Acoustic Parameter Set' (eGeMAPS) [14] low level descriptors (LLDs) extracted using our openSMILE toolkit [15]. This feature set is quite popular in the affective computing field already; we have used these exact features for establishing a baseline model performance for the AVEC 2017 challenge as the challenge organisers. The eGeMAPS LLDs is a minimalistic set of acoustic parameters, particularly tailor-made for affective vocalisation and voice research, consisting of only 23 LLDs. To capture the temporal dynamics of the individual parameters and LLD types, we extract BoAW features based on these LLDs. The BoAW approach generates a sparse fixed length histogram representation of the quantised features in time, thus not losing the temporal dynamics of the LLD vectors completely, all the while remaining noise-robust due to its inherent sparsity and the quantisation step [11, 16, 13].

However, the eGeMAPS LLDs are drastically different for the two databases in terms of their value ranges. Because the critical statistics – such as the mean, the variance, the maximum and the minimum value – are radically different (some with even the opposite signs), the statistics computed on one database cannot be reliably used to standardise or normalise the other database such that they share a common feature space. Furthermore, the codebook used in the AVEC 2017 challenge utilises a random sampling of the SEWA eGeMAPS LLD vectors. For transfer learning experiments however, we ideally should not generate the codebook by sampling only one of the two databases; a codebook that is likely to represent one dataset better. It is imperative to use an identical codebook to vector quantise the two databases that is completely data-independent – especially when the ranges of feature values are drastically different. It is only then that we can independently assess generalisability of the trained models ob-

jectively, free from effect of the codebook better representing temporal dynamics in one dataset over the other.

We thus generate a codebook of size 1000, independent of the two databases, consisting of 23-length LLDs. An array of shape 1000×23, populated with random samples from a normal distribution (mean= .5, standard deviation= .1) is used as a codebook matrix. We preprocess the LLDs by scaling and offsetting all of the data splits, using the offsets and the scaling factors that normalise the respective training split in the range $[0, 1]$. We then vector quantise all of the LLDs to the randomised codebook generated with 10 soft assignments for every LLD. We compute the distribution of the assignments in a moving window of 6 seconds, with a hop size of 0.1 seconds – similar to how AVEC 2017 features were generated [12].

3.3 Gold Standard Generation

We use the gold standard arousal and valence values of the AVEC 2017 challenge when training using the SEWA database [12]. We generate the gold standard for the GRAS^2 database using the same algorithm as of SEWA. The gold standard used in our previous studies on GRAS^2 differs only in that, we previously did not compensate for annotator-specific mean annotation standard deviations [11].

We use the modified *Evaluator Weighted Estimator* (EWE) method to generate the *gold standard*s, one per participant per emotion dimension. The goal of the *EWE* metric is to take into account the reliability of the individual annotators, signified by the weight r_k for every annotation y_k. This confidence value is computed by quantifying extent to which the annotations by that annotator agree with the rest of the annotations. The gold standard, y_{EWE} is defined as:

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

where $y_{n,k}$ is an 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. The symbol r_k is the corresponding annotator-specific weight. The lower bound for r_k is set to 0. In [17], the weight r_k is defined to be normalised cross-correlation between y_k and the averaged annotation sequence $\bar{y}_n$. The gold standards used in both the AVEC 2017 baseline paper [12] and the GRAS^2-based affect recognition study [11] redefined 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. The weights are lower bounded to 0 as usual. They are then normalised such that they sum to 1.

$$r''_{k_i,k_j} = \frac{\sum_{n=1}^{N} \left(y_{n,k_i} - \mu_{k_i}\right)\left(y_{n,k_j} - \mu_{k_j}\right)}{\sqrt{\sum_{n=1}^{N} \left(y_{n,k_i} - \mu_{k_i}\right)^2} \sqrt{\sum_{n=1}^{N} \left(y_{n,k_j} - \mu_{k_j}\right)^2}}, \text{ where: } \mu_k = \frac{1}{N} \sum_{n'=1}^{N} y_{n',k}, \tag{2}$$

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

3.4 Annotator Lag Compensation

To compensate for the reaction time of the annotators, we delay the feature vectors in time [18]. We use a delay value of 2.2 seconds, based on our previous grid search analysis on SEWA corpus [12]. In this study, we remove the repeating feature vectors at the beginning of every sample sequence introduced due to the lag compensating function used in AVEC 2017. We find that there is minute to no difference in performance because of removal of erroneously repeating feature vectors. This is expected, since the number of removed features (=22, in case of annotator lag compensation of 2.2 seconds) is less than 2 % of the total number of feature vectors for an average SEWA audio recording. Though it does not improve or deteriorate the performance of the models, we note this addition to our preprocessing steps in comparison with the AVEC 2017 workflow [12], for the sake of correctness and completeness.

3.5 Regression Models

For the new BoAW feature sets generated using a randomised codebook, we first generate baseline regression results by training support vector machine (SVM)-based regression models (SVR) using a linear kernel with complexity values, $C = [2^{-15}, 2^{-14}, ..., 2^{0}]$, just as was done when establishing the AVEC 2017 challenge baseline. We also experiment with additional C values in the range $[10^{-8}, ..., 10^{-5}]$ as the GRAS^2-trained arousal model was found to perform well for $C \in [2^{-15}, 2^{-7}]$. We ran regression models using simple feedforward neural networks (FFFN) and the double-stacked and a single-stacked recurrent neural network (RNN) with gated recurrent units (GRUs) in cascade with FFNNs. To train a GRU-based model, we used feature sequences of length 60, corresponding to 6 seconds. We experimented with several configurations for the network topologies (with 20 to 100 GRU nodes, 10 to 50-node layered FFNNs) , activation function permutations (selu, tanh, linear), feature lengths (60,80), learning rates (0.001 to 0.01 in the steps of 0.003), and optimisers (rmsprop, adam, adagrad, and adamax).

3.6 Post-processing

We post-process the predictions using the equation:

$$Y_{new} = (Y_{orig} - \mu_2)\frac{\sigma_1}{\sigma_2} + \mu_1, \tag{4}$$

where Y_{orig} is the primary prediction,Y_{new} is the post-processed prediction, μ_1, σ_1, μ_2, σ_2 are the mean and standard deviation of the training label sequence and the model's prediction on the training data respectively [19].

4 Results and Discussions

All of the models we trained (SVRs, GRU-RNNs and FFNNs) performed reasonably well, so long as the test split and the training splits came from the same

Table 2. Performance of the models in the transfer learning experiments for the arousal dimension. The models were trained only using the training split of the GRAS^2 database, and were tested on the remaining data splits of GRAS^2 and the entire SEWA German database. We note the performance on the individual data-splits of the SEWA database, to get better understanding of the coincidental data disparities and similarities between the two databases, and how the performance varies across splits with change in the complexity values. Interestingly enough, the similar SVR-based models trained on SEWA did not perform well on GRAS^2 database.

C Value	Database	Phase	Data split	CCC	PCC	RMSE
10^{-5}	GRAS^2	Training	Training	.501	.501	.137
		Validation	Validation	.363	.370	.144
		Testing	Testing	**.280**	.320	.152
	SEWA	Testing	Training	.171	.216	.149
			Validation	.325	.356	.144
			Testing	.197	.230	.132
			Entirety	**.223**	.263	.144
2^{-15}	GRAS^2	Training	Training	.582	.582	.125
		Validation	Validation	.382	.386	.140
		Testing	Testing	.266	.303	.149
	SEWA	Testing	Training	.128	.178	.170
			Validation	.280	.340	.161
			Testing	.188	.250	.144
			Entirety	.191	.252	.162
2^{-13}	GRAS^2	Training	Training	.691	.691	.108
		Validation	Validation	.350	.353	.144
		Testing	Testing	.241	.256	.143
	SEWA	Testing	Training	.082	.103	.188
			Validation	.236	.290	.184
			Testing	.155	.191	.160
			Entirety	.156	.193	.180
2^{-11}	GRAS^2	Training	Training	.778	.778	.091
		Validation	Validation	.331	.341	.152
		Testing	Testing	.228	.235	.144
	SEWA	Testing	Training	.107	.122	.198
			Validation	.251	.279	.195
			Testing	.171	.191	.169
			Entirety	.175	.196	.190
2^{-9}	GRAS^2	Training	Training	.834	.834	.079
		Validation	Validation	.248	.265	.170
		Testing	Testing	.180	.183	.145
	SEWA	Testing	Training	.120	.146	.233
			Validation	.156	.174	.231
			Testing	.208	.239	.186
			Entirety	.156	.181	.221

database, with concordance correlation coefficient (CCC) [20] close to 0.25 on an average. Of these, only the SVR-based models trained on GRAS^2 arousal annotations could reasonably make predictions in the transfer learning experiments (Table 2). The models otherwise mostly fail to generalise to a different dataset, with CCC values close to zero. For these transfer learning experiments from SEWA to GRAS^2, and vice versa, following are our key findings.

4.1 Neural networks tended to overfit to the primary database:

We observed the neural network-based models tended to overfit to the database they were trained on. The predictions were reasonably good for the test and validation splits of the same database that the training split came from. While performance on the same primary database depends also on the random initialisation of its weights and biases, the models invariably failed to make reasonable predictions on a different database (CCC close to zero).

4.2 Valence tracking learnings were not generalisable beyond the database:

A valence prediction is a particularly a harder problem as compared to an arousal prediction [11, 16, 3]. We observed that the models could predict the valence dimension for the validation and test splits of the same database (CCC as high as 0.42), but the prediction models tend to overfit to the database. This observation was irrespective of the type of model used, and the direction of transfer learning (i. e., whether SEWA to GRAS^2, or GRAS^2 to SEWA).

4.3 GRAS^2-trained SVR-based arousal tracking was reasonably generalised:

Interestingly though, an SVR-based arousal prediction models trained on GRAS^2 alone faired reasonably well on SEWA database with CCC values as high as 0.222 over the complete SEWA database – despite SEWA database being twice the size of GRAS^2. In the interest of reproducibility of the experiments presented in this paper, the complexity values and the corresponding performance values for the different models are as indicated in Table 2. We note that, out of the three SEWA splits, the model performs the worst on its training data split, which also is the most diversified split out of the three splits Table 1.

Despite having a lot smaller training set, the GRAS^2 to SEWA model transfer learning for the arousal prediction worked reasonably well. SEWA to GRAS^2 transfer learning however does not quite work (again, CCC close to zero), despite the training split having twice as much the data to train the model on, with an identical model parameters. We speculate that the SEWA database is not as in-the-wild as GRAS^2. GRAS^2 features also the random background speech, bustle, impact sounds, background music, and even the long non-speech sections. There exist emotion dimension labels for even these non-speech/rare-speech sections which the model needs to learn, which in itself is a challenging task. Such more in-the-wild nature of the data manifests itself in lot more challenging training instances that help model to learn arousal predictions with more nuances.

5 Conclusions and future work

We present a first-of-its-kind transfer learning study on the speech-based time-continuous in-the-wild affect recognising models. To this end, we used a novel BoAW approach that uses a novel data-independent randomised codebook. The GRAS^2 database – featuring relatively more observer's paradox-free affective behaviours, and a lot more data diversity in terms of conversation durations, acoustic events, noise dynamics, spontaneity of the featured affective behaviours – proved to be highly effective in training a more generalised arousal tracking model than the SEWA database, despite its smaller size. As for the valence dimension, none of the databases were effective enough in training a better-generalised valence tracking model. Furthermore, none of our neural network-based models could predict emotion dimensions (both arousal and valence) on a different database through transfer learning. All these models were observed to perform well on unseen data from the databases they were trained on.

The new BoAW paradigm of using the data-independent randomised codebooks helps one project dissimilar databases onto a common normalised feature space, while also inherently capturing the temporal dynamics of the LLDs; the technique which can be further developed and fine-tuned. We intend to investigate effect of different randomisation strategies (sampling from differently skewed distribution, or uniform or different normal distributions), also the codebook size and the number of assignments on the model performance.

We would like to also extend on this work by adding more in-the-wild databases. Our findings on better generalisability of the GRAS^2-trained arousal tracking model encourage us to use more of such databases that are free from the observer's paradox. Unfortunately, there are no other observer's paradox-free databases to work with, that are publicly available today. We plan to therefore collect new data using a similar data collection strategy used to build GRAS^2. The next logical step is to add other prominent affect recognition databases – such as RECOLA [3]. This will culminate into an exhaustive study on affect-related databases on their effectiveness in training the most-generalised, real-life time-continuous affect recognisers.

Acknowledgement


This work was partly supported by the EU's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA), and European Community's 7$^{\text{th}}$ Framework Program under the Grant No. 288587 (MASELTOV).

References

1. Paeschke, A., Kienast, M., Sendlmeier, W.F.: F0-contours in emotional speech. In: Proc. 14th Int. Congress of Phonetic Sciences. vol. 2, pp. 929–932 (1999)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Proc. 9th EUROSPEECH. pp. 1517–1520 (2005)
3. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA Multimodal Corpus of Remote Collaborative andAffective Interactions. In: 10th IEEE Int. Conf. Automat. Face and Gesture Recognition (FG'13). pp. 1–8 (2013)

4. Speer, S., Hutchby, I.: From Ethics to Analytics: Aspects of Participants' Orientations to the Presence and Relevance of Recording Devices. Sociology **37**(2), 315–337 (2003)
5. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective Multimodal Human-computer Interaction. In: Proc. 13th ACM MM. pp. 669–676. Singapore (2005)
6. Deng, J., Zhang, Z., Marchi, E., Schuller, B.: Sparse Autoencoder-based Feature Transfer Learning for Speech Emotion Recognition. In: Proc. 5th ACII. pp. 511–516. HUMAINE Association, IEEE, Geneva, Switzerland (2013)
7. Deng, J., Xia, R., Zhang, Z., Liu, Y., Schuller, B.: Introducing Shared-Hidden-Layer Autoencoders for Transfer Learning and their Application in Acoustic Emotion Recognition. In: Proc. 39th ICASSP. pp. 4851–4855. Florence, Italy (2014)
8. Coutinho, E., Deng, J., Schuller, B.: Transfer Learning Emotion Manifestation Across Music and Speech. In: Proc. IJCNN. pp. 3592–3598. Beijing, China (2014)
9. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proc. 17th ICMI. pp. 443–449. ACM (2015)
10. Eyben, F., Weninger, F., Paletta, L., Schuller, B.: 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. pp. 7–12. ACM, Sydney, Australia (2013)
11. Pandit, V., Cummins, N., Schmitt, M., Hantke, S., Graf, F., Paletta, L., Schuller, B.: Tracking Authentic and In-the-wild Emotions using Speech. In: Proc. 1st ACII Asia 2018. AAAC, IEEE, Beijing, P. R. China (2018)
12. Ringeval, F., Schuller, B., Valstar, M., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: 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. pp. 3–9. ACM, Mountain View, CA (2017)
13. Schmitt, M., Schuller, B.: openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. J. Mach. Learn. Res. **18** (2017)
14. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
15. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In: Proc. 21st ACM MM 2013. pp. 835–838. ACM, Barcelona, Spain (2013)
16. Schmitt, M., Ringeval, F., Schuller, B.: At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In: Proc. 17th INTERSPEECH . pp. 495–499. ISCA, San Francisco, CA (2016)
17. Grimm, M., Kroschel, K.: Evaluation of natural emotions using self assessment manikins. In: IEEE Workshop on Automat. Speech Recognition and Understanding (ASRU). pp. 381–385 (2005)
18. Mariooryad, S., Busso, C.: Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: Affective Computing and Intelligent Interaction (ACII). pp. 85–90. HUMAINE Association (2013)
19. Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B., Zafeiriou, S.: Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In: Proc. 41st ICASSP. pp. 5200–5204. IEEE, Shanghai, P. R. China (2016)
20. Lawrence, I., Lin, K.: A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics **45**(1), 255–268 (1989)