

EAT – The ICMI 2018 Eating Analysis and Tracking Challenge

Simone Hantke

Machine Intelligence & Signal Processing Group,
Technische Universität München, Germany
simone.hantke@tum.de

Panagiotis Tzirakis

GLAM – Group on Language, Audio & Music,
Imperial College London, UK

Maximilian Schmitt

ZD.B Chair of Embedded Intelligence for
Health Care and Wellbeing,
University of Augsburg, Germany

Björn Schuller

audEERING GmbH, Gilching, Germany
ZD.B Chair of Embedded Intelligence for
Health Care and Wellbeing,
University of Augsburg, Germany
GLAM – Group on Language, Audio & Music,
Imperial College London, UK

ABSTRACT

The multimodal recognition of eating condition – whether a person is eating or not – and if yes, which food type, is a new research domain in the area of speech and video processing that has many promising applications for future multimodal interfaces such as adapting speech recognition or lip reading systems to different eating conditions. We herein describe the ICMI 2018 EATING ANALYSIS AND TRACKING (EAT) CHALLENGE and address for the first time in research competitions under well-defined conditions new classification tasks in the area of user data analysis, namely audio-visual classifications of user eating conditions. We define three Sub-Challenges based on classification tasks in which participants are encouraged to use speech and/or video recordings of the audio-visual iHEARU-EAT database. In this paper, we describe the dataset, the Sub-Challenges, their conditions, and the baseline feature extraction and performance measures as provided to the participants.

CCS CONCEPTS

• **Information systems** → **Multimedia databases; Multimedia and multimodal retrieval; Task models;** • **Computing methodologies** → **Speech recognition;**

KEYWORDS

Challenge, Multimodal Data Analysis, Eating Condition, Human Behaviour

ACM Reference Format:

Simone Hantke, Maximilian Schmitt, Panagiotis Tzirakis, and Björn Schuller. 2018. EAT – The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of ACM International Conference on Multimodal Interaction (ICMI’18)*. ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ICMI’18, October 2018, Boulder, Colorado USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06...\$15.00

https://doi.org/10.475/123_4

1 INTRODUCTION

Within this Section, we will describe our first, open, audio-visual challenge, namely the ICMI 2018 EATING ANALYSIS AND TRACKING (EAT) CHALLENGE. Herein, we propose three Sub-Challenges, where participants may contribute based on classification tasks in which participants are encouraged to use speech and/or video recordings:

- (1) **FOOD-TYPE SUB-CHALLENGE:** Participants perform a seven-class food classification per utterance
- (2) **LIKABILITY SUB-CHALLENGE:** Recognise the subjects’ likability of each food type
- (3) **CHEWING AND SPEAKING DIFFICULTY SUB-CHALLENGE:** Recognise the level of difficulty to speak while eating.

Participants can perform one or more of these Sub-Challenges. For all Sub-Challenges, a target class label has to be predicted per clip, where each file contains one full speech utterance.

The challenge data itself contains audio, video and meta-data. The meta-data is composed of subject identity, age, and gender. The participants are welcome to use any combination of modalities. They can employ their own features and machine learning algorithms; however, a standard feature set is provided that may be used for both audio and video data. Baseline predictions as well as the baseline code in the three Sub-Challenges will be provided. Participants will have to stick to the pre-defined training/test splits and as these sets are subject independent, test data cannot be used for training purposes. Participants may report development results obtained from the training set (preferably with the supplied evaluation setups), but have only a limited number of three trials to upload their results on the test sets for the Sub-Challenges, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which undergoes a double-blind peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge.

As evaluation measures, for the FOOD-TYPE and LIKABILITY Sub-Challenges, we employ Unweighted Average Recall (UAR, i. e., the average of the class-specific recall values), especially because it is more adequate for (more or less unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) and has been

applied for the Eating Condition (EC) Sub-Challenge in the earlier organised INTERSPEECH 2015 Computational Paralinguistics Challenge (Interspeech 2015 ComParE) [18]. For the CHEWING AND SPEAKING DIFFICULTY Sub-Challenge, we make use of the Concordance Correlation Coefficient (CCC) [10], as it is a compromise between the mean squared error and the (linear) Pearson's Correlation Coefficient and is widely used to evaluate regression tasks [14]. In preliminary experiments on the iHEARU-EAT dataset, we found that, the results are more meaningful and consistent between the training set and the test set using CCC over rank correlation, even when evaluated on single subjects.

In the following section we introduce the challenge corpora, then in Section 3 we describe the baseline experiments and results, and we conclude the paper in Section 4.

2 DATASET

For the EAT CHALLENGE, the audio tracks of the audio-visual iHEARU-EAT database are used [5], which were partly featured as a Sub-Challenge of the Interspeech 2015 ComParE Challenge [18].

For the iHEARU-EAT database, 30 subjects were recorded in a quiet, low reverberant office room at the Technische Universität München. Out of those 30 speakers, 15 subjects are female and 15 male, with a mean age of 26.1 years (standard deviation: 2.7 years). 27 of the subjects are German native speakers; one is Chinese, one is Indian, and one has a Tunisian origin, but they have a close-to-native competence in German. Prior to the actual recording, subjects performed practice trials to familiarise themselves with the recording procedure. None of the subjects reported significant speech impediments.

It was decided for food classes with partly similar consistency (for instance, crisps and biscuits) and partly dissimilar consistency (for instance, nectarine vs crisps). Moreover, the food classes represent snacks which are likely to be encountered in practical scenarios and enable the subjects to speak while eating. In order to control for the amount of food being consumed, and in particular to encourage subjects to actually eat while speaking, an assistant provided the subjects with a serving of fixed size prior to the recording of each utterance. The assistant was sitting behind an opaque screen to ensure that the subjects felt unobserved while they were eating and speaking. The serving size was chosen such as to enable a significant effect on the subjects' speech. The subjects were advised not to eat food during the experiment if they are allergic to or they did not like to eat for any other reason.

Both read and spontaneous speech were recorded. For the read speech, the German version of the standard text "The North Wind and the Sun" ("Der Nordwind und die Sonne") was chosen, which is frequently used in phonetics, as it is phonetically balanced. It contains 108 words (71 distinct) with 172 syllables [4]. The subjects had to read the whole text with each sort of food. Spontaneous speech was elicited by prompting subjects to briefly comment on, e. g., their favourite travel destination, genre of music, or sports activity. A typical session of one subject lasted about one hour. After the recordings, the participants were asked to self-report on how much they like each sort of food they were eating during the experiment. This was achieved by setting a continuous slider to a value ranging between 0–*dislike extremely* and 1–*like extremely*.

Table 1: Statistics of the iHEARU-EAT database: Number of instances per class in the train/test split used for the Challenge.

#	Train	Test	Σ
No Food	140	blinded	blinded
Apple	140	blinded	blinded
Nectarine	133	blinded	blinded
Banana	140	blinded	blinded
Crisp	140	blinded	blinded
Biscuit	133	blinded	blinded
Gummi bear	119	blinded	blinded
Σ	945	blinded	blinded

Furthermore, they were asked to specify on a 5-point Likert scale the difficulties they encountered in eating each sort of food while speaking. For this Challenge, the likability labels have been mapped to two discrete values ('Neutral', 'Like') considering the range of the individual subjects' ratings. For the 'NoFood' eating condition, the label was set to 'Neutral'. 'Disliking' does not appear in the data as subjects did not eat the corresponding food in the first place.

Afterwards, the recordings were segmented into units resembling speaker turns (the six pre-defined sentences in the read text, or responses to the spontaneous speech prompts). The speaker turns were segmented manually, in order to remove parts with only 'eating noise', which could make the classification task too easy. By construction, 1/7 of the speaker turns contain spontaneous speech. Note that there is a slight difference in the amount of utterances per class, because some subjects chose not to eat all the types of food.

Conforming with the Interspeech 2015 ComParE Sub-Challenge, the data were split speaker-independently into a training set (20 speakers) and test set (10 speakers), stratified by age and gender. Classification is done per speaker turn. The resulting numbers of instances per class and set are shown in Table 1.

3 EXPERIMENTS AND RESULTS

3.1 Crossmodal Bag-of-Words

In our first baseline script, we provide a *crossmodal bag-of-words* (XBOW) [17] approach. This approach combines the well-known *bag-of-audio-words* (BoAW) and *bag-of-visual-words* (BoVW) representations of numeric descriptors. In those approaches, the low-level descriptors (LLDs), i. e., frame-level features extracted from an audio or video clip are assigned to template audio/visual 'words' and a term-frequency histogram is generated. The template assignment step can be seen as a *vector quantisation* by choosing always the template with the lowest Euclidean distance. It was found, however, that assigning not only the *closest*, but the N_a closest templates, can improve the BoAW method [16]. This is sometimes referred to as *multiple assignments*. Finally, *term-frequency weighting* of the histogram values is usually applied, e. g., by a logarithmic weighting or *inverse-document frequency* weighting [13].

The resulting – fixed-length – feature vector is then the input of a machine learning algorithm, in our case, a *support vector machine* (SVM). For the generation of the *codebook* of templates, we use a *random sampling* of the LLDs in the training set. As acoustic LLDs,

we use the COMPARE feature set that has been employed successfully in many paralinguistic speech recognition tasks during the last few years [19]. It contains *spectral*, *cepstral*, *prosodic*, and *voice quality* features, in total 130 descriptors, extracted from audio frames (20 ms-60 ms) with a hop size (shift) of 10 ms. As visual features, we use 68 *facial landmarks*. Facial landmarks describe the location of significant points in terms of their location within the image, usually in the two-dimensional pixel plane, even though there are approaches to estimate also the image depth of the landmarks [1]. These points specify the corners of the eyes, the eyebrows, the mouth and the location of the eye pupils and the tip of the nose. The suitability of the BoAW and the XBOW approaches has already been shown for the tasks of emotion recognition [14, 16], medical diagnosis [15], and sentiment analysis [2].

The whole processing chain can be reproduced with a script provided in the package of the EAT CHALLENGE. We utilise only published open-source tools, more specifically, we used OPENS-MILE [3] for acoustic feature extraction, OPENFACE [1] for facial landmark extraction, OPENXBOW [17] for the generation of XBOW, and SCIKIT-LEARN [12] for training and evaluating the classifiers.

The facial landmarks are normalised on frame-level to be always in the range between 0 and 1. In our initial experiments, we found that, using only the mouth-related facial landmarks provides better results and we experienced the tendency of overfitting when using all facial landmarks. The acoustic LLDs were standardised to zero mean and unit variance (estimating the parameters from the training partition) and split into 8 groups, according to their feature types, and the vector quantisation was done independently for each group. Besides the *number of assignments* per frame, the *codebook size*, i. e., the number of templates, is an important parameter. Results for this optimisation process are shown in Table 2. For each modality (audio/video), the codebooks size was optimised from 200 to 2000. It is important to note that, the codebook size applies to each of the feature groups, so, for the audio domain, the actual length of the final feature vector is 8 times larger. It was found beforehand that, for the visual domain, 10 assignments per frame were more suitable than only 1 assignment per frame, as used for the audio domain. As a post-processing step, the logarithm was taken from all term-frequencies in order to compress their range. Due to the relatively low amount of 20 speakers (10 female, 10 male) in the training set, the complexity of the SVM (with a linear kernel) was optimised using a *Leave-One-Speaker-Out Cross-Validation (LOSO-CV)* on the training set, i. e., in each fold, all instances of one speaker were kept out for evaluation. Finally, the SVM was trained on all speakers with the complexity $[1e^{-5}, 1e^{-4}, \dots, 1e^0]$ performing best in LOSO-CV.

3.2 End-to-end Learning

For our second approach we utilise Deep Neural Networks (DNNs), which have been successfully used in the affective computing domain [8], and more particularly for emotion recognition [21] [20]. Most popular architectures are the *Convolutional Neural Networks (CNNs)* [11], which are used to extract features, and the *Long Short-Term Memory (LSTM)* [7] models, which are a type of *Recurrent Neural Networks (RNN)* and can capture the dynamics in sequential data. For our purposes we apply DNNs in an end-to-end

Table 2: Results for the automatic recognition for LOSO-CV and the Test set with OPENXBOW (audio/visual) and different codebook sizes using SVM (linear kernel) as a classifier. The complexity was optimised on LOSO-CV for each modality, task, and codebook size. Audio features have been standardised, the number of assignments is 1. For the video features, the number of assignments is 10. A logarithmic term frequency weighting was used for both modalities. Results for FOOD TYPE and LIKABILITY are in terms of UAR, results for DIFFICULTY are in terms of CCC.

Audio	FOOD TYPE		LIKABILITY		DIFFICULTY	
	LOSO	Test	LOSO	Test	LOSO	Test
200	58.6 %	65.6 %	64.8 %	51.7 %	.439	.506
400	63.1 %	66.8 %	64.0 %	50.9 %	.432	.482
600	61.5 %	68.3 %	64.6 %	51.5 %	.455	.530
800	63.1 %	66.7 %	64.5 %	50.6 %	.446	.500
1000	63.7 %	67.1 %	64.7 %	49.8 %	.440	.503
1200	64.3 %	67.2 %	66.5 %	54.2 %	.451	.515
1400	63.5 %	68.2 %	65.7 %	52.2 %	.470	.506
1600	63.5 %	67.1 %	65.7 %	53.5 %	.466	.505
1800	63.8 %	66.7 %	65.4 %	53.0 %	.466	.510
2000	63.1 %	68.1 %	65.3 %	53.4 %	.469	.505
Video	LOSO	Test	LOSO	Test	LOSO	Test
200	26.4 %	25.5 %	55.6 %	56.1 %	.246	.252
400	27.1 %	27.7 %	54.6 %	59.9 %	.198	.250
600	27.4 %	27.3 %	55.4 %	53.8 %	.199	.250
800	26.8 %	27.2 %	54.4 %	59.1 %	.232	.272
1000	28.0 %	27.0 %	53.9 %	62.2 %	.244	.229
1200	27.4 %	25.9 %	55.9 %	58.3 %	.238	.246
1400	27.7 %	26.7 %	54.4 %	55.3 %	.230	.235
1600	27.8 %	27.9 %	54.2 %	56.0 %	.173	.236
1800	26.7 %	28.0 %	54.4 %	57.9 %	.181	.263
2000	26.0 %	27.2 %	53.8 %	57.3 %	.176	.268

manner, i. e., using *raw* input information. In many applications such as emotion recognition it is important to consider inputs from different modalities (e. g., audio, visual, or physiology) as each modality provides complementary information to the task. For our purposes, we experiment with both unimodal and multimodal input. To this end, we utilise the END2YOU toolkit [22] which provides capabilities for multimodal profiling by end-to-end deep learning. More particularly, the modalities we use are the audio, video, and audiovisual. In order to try to enhance the performance of the models, we incorporate to our models landmarks. A short description of each of the models is provided below.

Preprocessing. Before feeding the data to the models, they are pre-processed so as to speed up their training. For the visual frames we change the range of the pixels from $[0, 255]$ to $[0, 1]$ by dividing their intensity values with 255. For the audio part, we first downsample the signal from the original sampling rate of 44.1 kHz to 16 kHz and we form audio frames of length 33 ms (a 532-dimensional vector) to match the visual frame rate.

Audio Model. Our audio model is comprised by two layers. The first layer extracts features from the raw signal using CNNs with 40

filters of size 20 and a max-pooling layer of size 2. The second layer extracts higher abstraction features with a CNN with 40 filters of size 40 and a max-pooling layer of size 10. A 2-layer LSTM of 256 units is used on top to capture the contextual information in the data.

Visual Model. For the extraction of features from the visual domain we utilise residual networks and more particularly the ResNet-50 [6], which has been used extensively in the computer vision domain. A 2-layer LSTM of 256 units is used on top of the model to capture the contextual information in the data.

Audiovisual Model. Our audiovisual model is comprised by both the audio and visual models that are used to extract features from the audio and visual information, respectively. The extracted features are then concatenated to form a multimodal feature vector, which is subsequently passed to a 2-layer LSTM of 256 units.

Incorporating Landmarks. It is possible to incorporate landmarks to the aforementioned models. This is accomplished by concatenating the landmarks to the extracted features of each model, before passing them to the LSTM module.

Before training our models we extract a quarter of examples from the training set to create a validation set. The instances of each set are selected so that a subject belongs to one of them but not both. Our models are trained using the Adam optimiser [9], and a fixed learning rate of 10^{-3} throughout all experiments. The loss function used is the softmax cross entropy loss. Due to memory limitations the batch size varies in our experiments from 1 to 5.

3.3 Challenge Baselines

Table 3 shows the baseline results. For each task, the modality (audio/video) or the crossmodal system performing best on the test set has been selected as the official baseline (in bold). The BoAW approach outperforms the end-to-end audio-only approach in all three tasks. For the FOOD TYPE and the DIFFICULTY tasks, this is also the method performing best on the test set. For LIKABILITY, the BoVW provide a larger UAR on the test set. The results show that an early fusion of the audio and the visual domain (XBOW) does not increase the performance. However, the hyperparameters of the BoAW and BoVW have been tuned independently, so far. Moreover, exploiting both domains could lead to better results using a late fusion of the unimodal systems.

Furthermore, Table 3 outlines the best results obtained using end-to-end deep learning models. Note that the validation set contains 5 out of the total 20 participants and it can automatically be extracted from the training set using the provided baseline scripts. The audio model provides better results than the visual one in all of the challenges. This was expected as the ResNet-50 we use requires a large amount of data to be trained efficiently, which is not the case in our dataset. However, the visual model does provide performance gains when it is combined with the audio one. Incorporating landmarks in our models slightly benefited their performance, except for the difficulty challenge. The table depicts only the best results.

Results obtained using end-to-end deep learning do not outperform the results from the crossmodal bag-of-words method. The main reason for this is the limited number of training examples that do not allow to train efficiently the feature extractors (audio and visual). In addition, the length of the videos is high (200 to 400

Table 3: Baseline results for all tasks and modalities. The official baselines are highlighted. Results for FOOD TYPE and LIKABILITY are in terms of UAR, results for DIFFICULTY are in terms of CCC.

Modality	OPENXBOW		END2YOU	
	LOSO	Test	Dev	Test
Task 1 – FOOD TYPE				
Audio-only	64.3 %	67.2 %	35.2%	32.8%
Video-only	28.0 %	27.0 %	27.1%	24.5%
Audio+Video	63.9 %	67.0 %	34.8%	33.6%
Task 2 – LIKABILITY				
Audio-only	66.5 %	54.2 %	55.1%	53.7%
Video-only	55.9 %	58.3 %	52.9%	50.9%
Audio+Video	65.5 %	51.8 %	55.1%	54.2%
Task 3 – DIFFICULTY				
Audio-only	.470	.506	.342	.323
Video-only	.246	.252	.264	.220
Audio+Video	.481	.501	.345	.311

frames) which provides extra difficulties to train the RNN, which has to be fully unrolled before making the prediction.

4 CONCLUDING REMARKS

This first open, audio-visual challenge under strictly comparable conditions, namely the ICMI 2018 EATING ANALYSIS AND TRACKING (EAT) CHALLENGE, brings three new fields of research: (i) FOOD-TYPE SUB-CHALLENGE – perform seven-class food classification per utterance; (ii) LIKABILITY SUB-CHALLENGE – Recognise the persons' likability of each food type; and (iii) CHEWING AND SPEAKING DIFFICULTY SUB-CHALLENGE – Recognise the level of difficulty to speak while eating. Herein, we feature state-of-the-art end-to-end learning baselines and the popular OPENXBOW toolkit.

Yet, feature sets and learning procedures are standard but not optimised and kept generic across the tasks, despite their obvious differences. For all computation steps, baseline and evaluation scripts are provided that can, but do not need to be used by the participants. We expect participants to obtain considerably better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular Sub-Challenges.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the European Community's Seventh Framework Programme through ERC Starting Grant No. 338164 (iHEARu) and the EPSRC Center for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1). We thank the sponsor of the Challenge: audEERING GmbH. Responsibility lies with the authors.

REFERENCES

- [1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 2016. OpenFace: an Open Source Facial Behavior Analysis Toolkit. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV/CA, 1–10.
- [2] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt, and B. Schuller. 2018. Multimodal Bag-of-Words for Cross Domains Sentiment Analysis. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Calgary, Canada, 1–5.
- [3] F. Eyben, F. Weninger, F. Groß, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. of ACM Multimedia*. Barcelona, Spain, 835–838.
- [4] Tino Haderlein, Cornelia Moers, Bernd Möbius, Frank Rosanowski, and Elmar Nöth. 2011. Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation. In *Proc. of Text, Speech and Dialogue (TSD)*. Berlin, Heidelberg, 195–202.
- [5] Simone Hantke, Felix Weninger, Richard Kurlle, Fabien Ringeval, Anton Batliner, Amr El-Desoky Mousa, and Björn Schuller. 2016. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Types, Use-Cases, and Impact on ASR Performance. *PLoS ONE* 11 (2016), 1–24.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA, 770–778.
- [7] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9 (1997), 1735–1780.
- [8] Y. Kim, H. Lee, and E. M. Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP*. 3687–3691.
- [9] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [10] I. Lawrence and K. Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* (1989), 255–268.
- [11] Yann LeCun et al. 1989. Generalization and network design strategies. *Connectionism in Perspective* (1989), 143–155.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [13] M. Riley, E. Heinen, and J. Ghosh. 2008. A Text Retrieval Approach to Content-based Audio Retrieval. In *Proc. International Symposium on Music Information Retrieval (ISMIR)*. Philadelphia, PA, 295–300.
- [14] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proc. of Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, CA, 3–9.
- [15] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller. 2016. A Bag-of-Audio-Words Approach for Snore Sounds – Excitation Localisation. In *Proc. of ITG Conference on Speech Communication*. Paderborn, Germany, 264–268.
- [16] M. Schmitt, F. Ringeval, and B. Schuller. 2016. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In *Proc. of INTERSPEECH*. San Francisco, CA, 495–499.
- [17] M. Schmitt and B. Schuller. 2017. openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. *Journal of Machine Learning Research* 18 (2017), 1–5.
- [18] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Höning, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. 2015. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition. In *Proc. of INTERSPEECH*. Dresden, Germany, 478–482.
- [19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. of INTERSPEECH*. Lyon, France, 148–152.
- [20] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP*. 5200–5204.
- [21] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. 2017. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing* 11 (2017), 1301–1309.
- [22] P. Tzirakis, S. Zafeiriou, and B. W. Schuller. 2018. End2You – The Imperial Toolkit for Multimodal Profiling by End-to-End Learning. *arXiv preprint arXiv:1802.01115* (2018).