Performance Analysis of Unimodal and Multimodal Models in Valence-Based Empathy Recognition

Adria Mallol-Ragolta[1], Maximilian Schmitt[1], Alice Baird[1],
Nicholas Cummins[1] and Björn Schuller[1,2]
[1] ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] GLAM – Group on Language, Audio & Music, Imperial College London, UK
adria.mallol-ragolta@informatik.uni-augsburg.de

Abstract— **The human ability to empathise is a core aspect of successful interpersonal relationships. In this regard, human-robot interaction can be improved through the automatic perception of empathy, among other human attributes, allowing robots to affectively adapt their actions to interactants' feelings in any given situation. This paper presents our contribution to the generalised track of the One-Minute Gradual (OMG) Empathy Prediction Challenge by describing our approach to predict a listener's valence during semi-scripted actor-listener interactions. We extract visual and acoustic features from the interactions and feed them into a bidirectional long short-term memory network to capture the time-dependencies of the valence-based empathy during the interactions. Generalised and personalised unimodal and multimodal valence-based empathy models are then trained to assess the impact of each modality on the system performance. Furthermore, we analyse if intra-subject dependencies on empathy perception affect the system performance. We assess the models by computing the concordance correlation coefficient (CCC) between the predicted and self-annotated valence scores. The results support the suitability of employing multimodal data to recognise participants' valence-based empathy during the interactions, and highlight the subject-dependency of empathy. In particular, we obtained our best result with a personalised multimodal model, which achieved a CCC of 0.11 on the test set.**

I. INTRODUCTION

Empathy is the ability to recognise and relate to the emotional needs of others, and can be an important human trait in improving interpersonal relationships [19]. In the field of medicine, empathy towards the patient is a vital part of the care-giving process [16]. In psychotherapy, for instance, treatments from therapists with high empathy towards their patients have shown to be more effective compared to those from therapists with low empathy [18], [21]. Despite this, the ability to empathise is often overlooked by medical professionals in favour of technological advancements in diagnosis and treatment [36].

Researchers in the field of human-robot interaction (HRI) have envisioned the use of automatic empathy prediction systems to improve HRI [13], [5]. In this regard, previous approaches analysing empathy from audio- and video-based features have resulted in promising findings [39] due to the richness in emotional content of both modalities. With this in mind, research efforts towards empathy prediction have explored the use of audio alone [38], in part, due to the success achieved in the field of computational paralinguistics of speech [30]. Nevertheless, as it has been shown that the modelling of facial information is vital in the context of empathy [33], in recent years, contributions towards empathy prediction have also come from the domain of computer vision [15], [7].

This paper reflects our contribution to the generalised track of the One-Minute Gradual (OMG) Empathy Prediction Challenge[1]. Our approach explores the use of unimodal and multimodal models on the proposed task with audio- and video-based features. Our main hypothesis relies on the suitability of multimodal models for automatic valence-based empathy recognition, since the combination of acoustic and visual information should be more effective than the utilisation of one modality alone, as suggested in previous emotion detection research [37], [25]. We employ OPENSMILE [12] to extract the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [11] from the speech signals, since it has been very successful in similar affective computing tasks, such as depression recognition [32], [35]. As visual features, we extract facial action units (FAUs) using OPENFACE [3], [6], because FAUs have been shown to be an effective method for recognising a variety of emotional states [9], [28], [27]. These features are then employed to train unimodal and multimodal recurrent models based on bidirectional long short-term memory networks (BLSTM) [14] to capture valence-based empathy time-dependencies.

The rest of the paper is laid out as follows. Section II introduces the dataset utilised, while Section III describes the methodology followed. Section IV presents the results obtained from the experiments performed, while Section V concludes the paper and points out some future work directions.

II. OMG-EMPATHY PREDICTION DATASET

The OMG Empathy Prediction Dataset used in this work has been provided by the OMG-Empathy Prediction Challenge organisers. This corpus contains dyadic interactions between 4 different actors and 10 different listeners talking to each other about 8 predefined topics related to one or more emotional states. Actors were instructed to maintain control

[1] https://www2.informatik.uni-hamburg.de/wtm/omgchallenges/omg_empathy.html

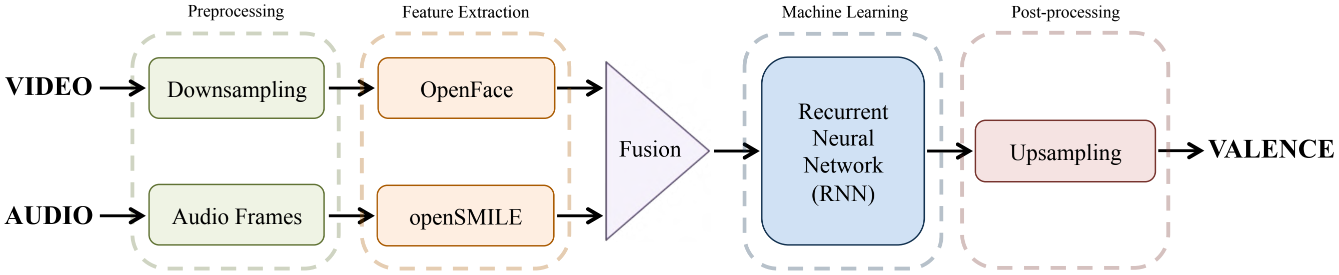

Fig. 1: Block diagram illustrating the preprocessing, feature extraction, machine learning and post-processing stages of the valence-based empathy recognition system implemented and the workflow of the audio and video signals employed as input.

TABLE I: OMG-Empathy Prediction Dataset summary with the total number of interactions recorded, the total number of frames available, and the number of frames used in our approach after the preprocessing stage.

	Train. Set	Devel. Set	Test Set	$\sum$
Interactions	40	10	30	80
Original frames	309 875	95 575	228 400	633 850
Used frames	61 975	19 115	45 680	126 770

over semi-scripted conversations, although improvisation was encouraged, so a natural conversation scenario could be recorded.

After each interaction, listeners watched their own recording on a computer screen and self-assessed their feelings in terms of valence on a continuous scale ranging from positive to negative values with a joystick. In order to collect different listeners' reaction levels for the same actor, each actor was assigned to take part in the interactions corresponding to two specific topics. Hence, a total of 80 interactions – equivalent to 422 minutes and 34 seconds of video – were recorded, which were split into 40, 10 and 30 instances for training, development and test sets, respectively (cf. Table I). Audio-visual data was recorded from both the actor and the listener during the interaction: while audio signals were recorded at a sampling rate (f_s) of 16 kHz, video data was recorded at a f_s of 25 fps. Valence annotations were continuously collected at a f_s of 25 Hz.

III. METHODOLOGY

In this work, we implement a valence-based empathy recognition system that receives audio-visual data recorded from a conversation between an actor and a listener and predicts the listener's valence throughout the interaction. In a real scenario, such a system might be used by a non-human agent to recognise the empathy of its interaction partner. Thus, our approach utilises visual information from the listener and acoustic information from both the actor and the listener, since paralinguistics in actors' utterances might also impact on listeners' empathy. The way the actor's utterances *are being said* can affect the listener's perception of the actor and, as a consequence, interfere in the interaction. This section presents the implemented system (cf. Figure 1), which is publicly available[2], and describes its main stages.

A. Preprocessing

The first stage of the system requires a procedure to overcome the disparity between audio and video sampling rates, as there are 640 audio samples per video frame. Previous works concerning the automatic prediction of affective information from multimodal data employed different solutions to overcome the disparity in sampling rates [31], [24]. In this regard, our approach adopts the segmentation of the original audio signals into audio frames. This technique is based on selecting portions of the original audio stream that contain those samples corresponding to the temporal length equivalent to one frame of video.

Taking into account the number of video frames available in the dataset, and with the aim of speeding up the training time of the valence-based empathy models, we reduce the temporal resolution of the listeners' videos by selecting 1 in 5 consecutive video frames. As a result, the video sampling rate is lowered to 5 fps and so it is the sampling rate of the valence annotations (in this case, to 5 Hz) to synchronise both modalities. We expect this downsampling to have a small impact in the overall system performance, as valence annotations are stable with neither marked nor fast changes over time. Furthermore, we accept the information loss caused by this sampling rate, as lower sampling rates were used to re-sample visual data in the affective computing literature [31].

In order not to reduce the acoustic information but still meet the matching requirement with the new sampling rate, our approach generates audio frames equivalent to five video frames. Additionally, this audio segmentation is performed in such a way that consecutive audio frames overlap. In particular, each audio frame contains 50 % of the samples from the preceding and the proceeding audio frames. Thus, each audio frame has a length of 0.4 seconds. Zero padding is used in the first and last audio frame, when necessary, by adding zeros to the samples of the audio frames corresponding to the non-existent preceding or proceeding audio frames.

B. Feature Extraction

The next step in the pipeline of the valence-based empathy recognition system corresponds to the extraction of features from the input modalities. Due to the relevance of facial information when judging behavioural cues [1], we hypothesise that listeners' faces might contain information related to their level of empathy. Thus, we extract 35 visual features

[2] `https://github.com/EIHW/OMGempathy2019`

from the listener's face using the OpenFace software [3], which captures the intensity of 17 FAUs and the presence of 18 FAUs. The values of the intensity features range from 0.00 to 5.00: 0.00 indicates the absence of a particular FAU, while 1.00 and 5.00 indicate the presence of a particular FAU at minimum and maximum intensities, respectively. On the other hand, the values of the presence features are 0 or 1, indicating the absence or presence of a particular FAU. The acoustic features that we extract from the segmented audio frames correspond to the 88 features defined by the eGeMAPS feature set [11]. As outlined, eGeMAPS is a feature set widely used in the affective computing literature [34], [26], [29], [10], and it is extracted using the open-source openSMILE software [12].

For every interaction in the dataset, we extract visual features $\in \mathbb{R}^{N \times 35}$ and acoustic features $\in \mathbb{R}^{N \times 88}$, where N corresponds to the total number of frames available from each interaction to train our valence-based empathy models. Additionally, we would like to point out that, although no normalisation is performed on the visual features because of their bounded range, acoustic features are z-normalised [8], so they are zero-mean and unit-variance.

C. Machine Learning

The time-dependencies of changes in empathy throughout the interactions are modelled using a recurrent neural network (RNN), in particular a bidirectional long short-term memory network (BLSTM). The architecture of the neural network we propose employs a BLSTM layer with $M \in [30, 40, 50]$ LSTM units, which is optimised on the development set, followed by a dense layer with a single unit, which outputs the valence predictions at every time step. Respectively, *tanh* and *linear* activation functions are employed in both layers. The parameters of the neural network are optimised with the concordance correlation coefficient (CCC) as the loss function with *Adam* as the optimiser. From preliminary experiments, the batch size is fixed to five, and the number of epochs with no improvement allowed before stopping training is fixed to three.

D. Post-processing

The last stage of the pipeline requires post-processing the valence predictions to revert the downsampling performed on the interaction videos at the preprocessing stage. First, we use a median filter with a kernel size of 301 samples to remove possible noise from the predicted annotations. Next, considering the nature of the ground truth valence annotations over time, we decided to use replication as the upsampling method. Therefore, each predicted annotation at low temporal resolution is replicated five times to reach the sampling rate of the original annotations.

IV. EXPERIMENTAL RESULTS

The main goal of this work is to analyse the performance differences between unimodal and multimodal models for valence-based empathy recognition systems. While unimodal models are trained with visual or acoustic features alone, feature-level fusion is employed to train multimodal models by means of concatenating the features from both modalities.

TABLE II: Comparison of the concordance correlation coefficients (CCC) computed on the development set of the OMG-Empathy Prediction Dataset with generalised and personalised unimodal and multimodal models. The 3 best results obtained are highlighted.

	Generalised models			Personalised models		
Acoustic model	0.04	-0.00	0.03	0.02	0.04	0.05
Visual model	-0.02	-0.02	-0.01	-0.02	-0.01	0.04
Multimodal model	0.05	**0.07**	**0.06**	-0.02	0.01	**0.06**
LSTM units	30	40	50	30	40	50

The first approach to solve the generalised track of the challenge is based on the use of generalised models, which are trained with all of the interactions available for training, regardless of the listeners' identities. Analysing the self-assessed annotations, however, we observe that the labelled valence scores seem to be listener dependent. As generalised models might not be able to capture intra-subject dependencies on the perception of empathy [20], we also train personalised models, which are trained exclusively with interactions corresponding to a particular listener, to solve the same task. Furthermore, we expect these personalised models to capture the listeners' dependencies on the self-perception of empathy. This way, performance differences, if any, between generalised and personalised models can be compared.

Generalised and personalised unimodal and multimodal models are evaluated by computing the CCC between the self-assessed annotations and valence scores predicted with the trained models. Results computed on the development set for generalised and personalised unimodal and multimodal models with LSTM units $\in [30, 40, 50]$ are presented in Table II.

The results computed from the generalised models highlight the suitability of employing multimodal models for the recognition of empathy, as they provide greater CCC scores than unimodal models, irrespective of the number of LSTM units used. For instance, when using 40 LSTM units we obtain $CCC_{multimodal} = 0.07$, $CCC_{visual} = -0.02$ and $CCC_{acoustic} = -0.00$, while when using 50 LSTM units we achieve $CCC_{multimodal} = 0.06$, $CCC_{visual} = -0.01$ and $CCC_{acoustic} = 0.03$. On the other hand, the results obtained from the unimodal models suggest that the visual features extracted alone are not a suitable predictor of empathy, as they obtain the lowest CCC scores. As visual data was successfully used on valence prediction problems [22], this is a surprising result, which allows us to argue about the suitability of the visual feature set employed in this work for the automatic recognition of empathy.

Analysing the results measured from the personalised models, we observe that the best CCC scores are obtained when the BLSTM layer employs 50 LSTM units. The results obtained for this particular network configuration are $CCC_{multimodal} = 0.06$, $CCC_{visual} = 0.04$ and $CCC_{acoustic} = 0.05$. In this case, we also observe

TABLE III: Comparison of the concordance correlation coefficients (CCC) computed on the development and testing sets of the OMG-Empathy Prediction Dataset with the selected models to participate in the challenge. The best result obtained on the testing set is highlighted.

	Devel. Set	Test Set
Generalised multimodal model (40 LSTM units)	0.07	0.05
Generalised multimodal model (50 LSTM units)	0.06	0.06
Personalised multimodal model (50 LSTM units)	0.06	**0.11**

that multimodal models obtain a higher CCC score, in comparison to the unimodal models, and that the lowest CCC score is obtained from the visual models.

In summary, we highlighted that personalised unimodal models with 50 LSTM units performed better than generalised unimodal models. Furthermore, personalised multimodal models with 50 LSTM units performed at par with generalised multimodal models. Hence, these results support the suitability of employing personalised models to capture the intra-subject dependencies on the perception of empathy, although each personalised model is only tested on a single instance due to the limited data available in the development set, which complicates the draw of a valid conclusion. Finally, we need to state that none of our results reach the baseline provided by the challenge organisers, which was set to $CCC = 0.111$ for the generalised track. This result casts doubt on the generalisation capabilities of the trained models, which might be due to an insufficient amount of training data, or the reduced representation of the corpus in the development set.

Participants of the challenge have three attempts to predict valence annotations on the test set. In this regard, the three models we select are the generalised multimodal models with 40 and 50 LSTM units, and the personalised multimodal models with 50 LSTM units, as they provide the highest CCC scores on the development set. At this point, we would like to note that during the development phase of our system, models are trained on interactions belonging exclusively to the training set, and tested on the development set. When the development phase is completed, models are trained on interactions from both training and development sets, and tested on the test set. Having said this, the results obtained on the test set with the selected models are summarised in Table III.

From the results obtained on the test set, we observe that generalised models perform at par in both development and test sets, which lowers our confidence in the suitability of generalised models for automatic valence-based empathy recognition problems. Nonetheless, for the personalised models, we observe a performance improvement on the test set, $CCC_{multimodal} = 0.11$, with respect to the development set, $CCC_{multimodal} = 0.06$. This result supports the suitability of employing personalised models to automatically predict empathy, and it also suggests the existence of intra-subject dependencies on the perception of empathy, which cannot be captured with generalised models.

Regardless of the scenario, the CCC scores measured are low, which suggests that we might be modelling noise. Furthermore, the median filter used in the post-processing stage might impact the system performance, as it smooths the predictions. Training the models with windowed shifted portions of the interactions could help improve system performance, as we will increase the size of the training data.

V. Conclusions and Future Work

In this paper, we presented our approach to the generalised track of the OMG-Empathy Prediction Challenge with the aim of analysing the impact of unimodal and multimodal models on the automatic recognition of valence-based empathy. The results obtained on the development set supported our hypothesis that both acoustic and visual signals provided relevant information for empathy modelling, as multimodal models achieved better CCC scores than unimodal models. Furthermore, the stronger performances exhibited by our personalised models indicate the existence of intra-subject dependencies on the perception of empathy. We obtained our best experimental results using personalised multimodal models that use 50 LSTM units in the bidirectional layer, which resulted in a CCC of 0.06 on the development set and a CCC of 0.11 on the test set.

In future work, we plan to explore the benefits of realigning the audio-visual and self-assessed information using annotation delay compensation [17]. Speaker diarisation [4] could also be a conducive next step, concerning the individual impact of utterances from actors or listeners on the performance of valence-based empathy recognition systems. We also plan to assess the benefits of alternative feature spaces [23], [2], in addition to the further investigation of personalised models employing deeper neural network architectures with additional hidden layers that would help to improve the system performance. Furthermore, we will also consider the use of transfer learning methods to provide more training data and, at the same time, analyse the task-dependency of valence-based empathy models.

VI. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu). Additionally, this project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), and from the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation Bavaria (ZD.B).

References

[1] N. Ambady and R. Rosenthal. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[2] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller. Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio. In *Proceedings of Detection and Classification of Acoustic Scenes and Events*, 5 pages, Munich, Germany, September 2017.

[3] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. OpenFace: A General-purpose Face Recognition Library with Mobile Applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, February 2012.

[5] M. Anshar and M.-A. Williams. Evolving Robot Empathy towards Humans with Motor Disabilities through Artificial Pain Generation. *AIMS Neuroscience*, 5(1):56–73, 2018.

[6] T. Baltrušaitis, P. Robinson, and L. Morency. OpenFace: an Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 10 pages, Lake Placid, NY, USA, March 2016. IEEE.

[7] N. Churamani, P. Barros, E. Strahl, and S. Wermter. Learning Empathy-Driven Emotion Expressions using Affective Modulations. In *Proceedings of the International Joint Conference on Neural Networks*, 8 pages, Rio de Janeiro, Brazil, 2018. IEEE.

[8] D. Clark-Carter. z scores. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, editors, *Wiley StatsRef: Statistics Reference Online*. 2014.

[9] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre. Detecting Depression from Facial Actions and Vocal Prosody. In *Proceedings of the 3^{rd} International Conference on Affective Computing and Intelligent Interaction and Workshops*, 7 pages, Amsterdam, Netherlands, September 2009.

[10] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25^{th} ACM International Conference on Multimedia*, pages 478–484, Mountain View, CA, USA, 2017. ACM.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016.

[12] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the 18^{th} ACM International Conference on Multimedia*, pages 1459–1462, Firenze, Italy, 2010. ACM.

[13] P. Fung, D. Bertero, Y. Wan, A. Dey, R. H. Y. Chan, F. B. Siddique, Y. Yang, C.-S. Wu, and R. Lin. Towards Empathetic Human-Robot Interactions. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 173–193, Konya, Turkey, 2016. Springer.

[14] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016.

[15] D. Greenwood, S. Laycock, and I. Matthews. Predicting Head Pose in Dyadic Conversation. In *Proceedings of the International Conference on Intelligent Virtual Agents*, pages 160–169, Stockholm, Sweden, 2017. Springer.

[16] M. Hojat, S. Mangione, T. J. Nasca, M. J. Cohen, J. S. Gonnella, J. B. Erdmann, J. Veloski, and M. Magee. The Jefferson Scale of Physician Empathy: Development and Preliminary Psychometric Data. *Educational and psychological measurement*, 61(2):349–365, 2001.

[17] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps. An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. In *Proceedings of the 5^{th} International Workshop on Audio/Visual Emotion Challenge*, pages 41–48, Brisbane, Australia, 2015. ACM.

[18] B. D. Jani, D. N. Blane, and S. W. Mercer. The Role of Empathy in Therapy and the Physician-Patient Relationship. *Complementary Medicine Research*, 19(5):252–257, 2012.

[19] M. J. Lambert and A. E. Bergin. The Effectiveness of Psychotherapy. *Handbook of psychotherapy and behavior change*, 4:143–189, 1994.

[20] R. Lennon and N. Eisenberg. Gender and Age Differences in Empathy and Sympathy. In N. Eisenberg and J. Strayer, editors, *Empathy and Its Development*, chapter 9, pages 195–217. Cambridge University Press, 1990.

[21] T. B. Moyers and W. R. Miller. Is Low Therapist Empathy Toxic? *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, 27(3):878–884, 2013.

[22] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing*, 2(2):92–105, April 2011.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In M. W. J. Xianghua Xie and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference*, pages 41.1–41.12, Swansea, UK, September 2015. BMVA Press.

[24] H. Rao, Z. Ye, Y. Li, M. A. Clements, A. Rozga, and J. M. Rehg. Combining Acoustic and Visual Features to Detect Laughter in Adults' Speech. In *Proceedings of the 1^{st} Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing*, pages 153–156, Vienna, Austria, 2015.

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

[26] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller. Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children. In *Proceedings of Interspeech*, pages 1210–1214, San Francisco, CA, US, 2016.

[27] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. A. Salah, and M. Pantic. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13, Seoul, Republic of Korea, 2018. ACM.

[28] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7^{th} Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, Mountain View, CA, USA, 2017. ACM.

[29] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller. Towards Cross-lingual Automatic Diagnosis of Autism Spectrum Condition in Children's Voices. In *Speech Communication; 12. ITG Symposium*, pages 264–268, Paderborn, Germany, October 2016. VDE.

[30] B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[31] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, January-March 2016.

[32] B. Stasak, J. Epps, N. Cummins, and R. Goecke. An Investigation of Emotional Speech in Depression Classification. In *Proceedings of Interspeech*, pages 485–489, San Francisco, CA, USA, 2016.

[33] J. Sun and P. Liu. The Advantage of Empathy from Evidence of Visual Processing. *Advances in Psychological Science*, 25(suppl.):7, 2017.

[34] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5200–5204, Shanghai, China, March 2016. IEEE.

[35] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller. Implementing Gender-Dependent Vowel-Level Analysis for Boosting Speech-Based Depression Recognition. In *Proceedings of Interspeech*, pages 3266–3270, Stockholm, Sweden, August 2017. ISCA.

[36] M. Weiner and P. Biondich. The Influence of Information Technology on Patient-Physician Relationships. *Journal of General Internal Medicine*, 21(1):35–39, 2006.

[37] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan. Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling. In *Proceedings of Interspeech*, pages 2362–2365, Makuhari, Chiba, Japan, 2010. ISCA.

[38] B. Xiao, C. Huang, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan. A Technology Prototype System for Rating Therapist Empathy from Audio Recordings in Addiction Counseling. *PeerJ Computer Science*, 2:e59, 2016.

[39] B. Xiao, Z. E. Imel, P. Georgiou, D. C. Atkins, and S. S. Narayanan. Computational Analysis and Simulation of Empathic Behaviors: a Survey of Empathy Modeling with Behavioral Signal Processing Framework. *Current Psychiatry Reports*, 18(5):49, 2016.