An Investigation of Cross-Cultural Semi-Supervised Learning for Continuous Affect Recognition

Adria Mallol-Ragolta[1], *Nicholas Cummins*[1,2], *and Björn W. Schuller*[1,3]

[1] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Department of Biostatistics and Health Informatics, IoPPN, King's College London, UK
[3] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

adria.mallol-ragolta@informatik.uni-augsburg.de

Abstract

One of the keys for supervised learning techniques to succeed resides in the access to vast amounts of labelled training data. The process of data collection, however, is expensive, time-consuming, and application dependent. In the current digital era, data can be collected continuously. This continuity renders data annotation into an endless task, which potentially, in problems such as emotion recognition, requires annotators with different cultural backgrounds. Herein, we study the impact of utilising data from different cultures in a semi-supervised learning approach to label training material for the automatic recognition of arousal and valence. Specifically, we compare the performance of culture-specific affect recognition models trained with manual or cross-cultural automatic annotations. The experiments performed in this work use the dataset released for the Cross-cultural Emotion Sub-challenge of the Audio/Visual Emotion Challenge (AVEC) 2019. The results obtained convey that the cultures used for training impact on the system performance. Furthermore, in most of the scenarios assessed, affect recognition models trained with hybrid solutions, combining manual and automatic annotations, surpass the baseline model, which was exclusively trained with manual annotations.

Index Terms: continuous affect recognition, cross-cultural analysis, audiovisual processing, semi-supervised learning.

1. Introduction

High quality labelled data is of vital importance in supervised learning approaches. The increasing amount of sensors and devices permanently connected to the Internet allows the continuous collection of information. So that this data can help improving the performance of machine learning algorithms, it needs to be annotated. Data collection can, therefore, be expensive and time-consuming. This process is even costlier when it comes to affective datasets, as the gold standard being mapped to a specific sample is determined by analysing the individual labels provided by multiple annotators on the same sample. Furthermore, these annotators need specific training, and are culturally specific. The annotators should share the same culture as the users in the dataset to guarantee annotation reliability, as different cultures show emotions differently [1, 2]. To ease the data annotation process, researchers have investigated the use of *Semi-Supervised Learning* (SSL) approaches [3].

Methods using SSL have been investigated with different modalities [4, 5, 6, 7, 8, 9]. In the particular case of the audio modality, SSL techniques have been employed in a wide range of problems, such as automatic speech recognition [10], sound classification [11], or depression detection [12], to name but a few. In the field of affective computing, researchers investigated the benefits of SSL in the problem of emotion recognition from audio [13] and video [14, 15]. Previous works proposed methods to enhance the annotations inferred via SSL to mitigate the propagation of the error caused by the inference, reducing their impact to the overall system performance [16]. Further studies explored cooperative [17] and collaborative [18] learning approaches, which combine expert (manual) and machine (automatic) annotations. Others investigated the benefits of using SSL in crowdsourcing paradigms to generate emotional labels [19].

The possibility to automatically annotate affective datasets, or to reduce the number of annotators needed for labelling without deteriorating the quality of the annotations themselves is the primary goal when using SSL. Despite the usefulness of SSL techniques in affect-related problems, to the best of the authors' knowledge, the limitations of SSL in cross-cultural settings has not been investigated yet. SSL-powered systems that automatically gather data from online social media platforms [20] might benefit from these investigations in order to determine whether cultural aspects need to be taken into account for improving the quality of their annotations. In this work, we aim to analyse how the cultural dependencies on conveying emotions impact the performance of affect recognition models when using SSL annotations as training material. Specifically, we focus this study on the continuous recognition of arousal from the voice, and valence from the face, assessing our models on German, Hungarian, and Chinese cultures.

The rest of the paper is laid out as follows. Section 2 presents the dataset employed, while Section 3 describes the methodology followed. Section 4 details the experiments performed and analyses the results obtained. Finally, Section 5 concludes the paper and suggests some future work directions.

2. Cross-cultural Emotion Dataset

The present work investigates the *Cross-cultural Emotion Sub-challenge* (CES) dataset, an audio-visual dataset with continuous emotional annotations in the valence-arousal space [21]. The dataset was released for the CES task of the 9^{th} *Audio/Visual Emotion Challenge* (AVEC) and Workshop [22], and consists of a subset of the interactions gathered in the SEWA database [23]. CES captures spontaneous in-the-wild interactions between pairs of friends or relatives from German, Hungarian, and Chinese cultures, while remotely discussing a commercial they had just seen. The German and Hungarian cultures were available in the train, development, and test partitions. Interactions from the Chinese culture were only available in the test partition.

The interactions were recorded using a computer-based platform. Audio data was recorded at 48 kHz, video data at 50 *Frames Per Second* (FPS), and affect-related annotations at 10 Hz. The video modality always contains information from one of the two interactants, while the audio modality contains

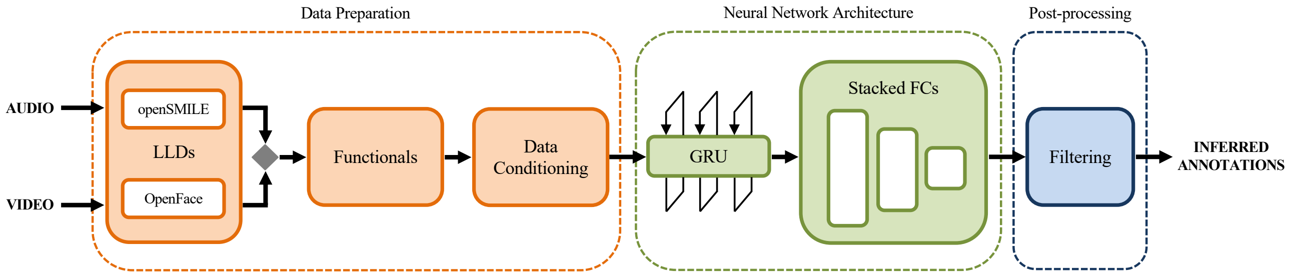

Figure 1: *Block diagram illustrating our system. Audio or video signals are received as input, and Low Level Descriptors (LLDs) are extracted from them using* OPENSMILE *or* OPENFACE*, respectively. Functionals are computed from the LLDs, and arranged in fixed-length sequences. These are then fed into a single-layer Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) followed by three stacked Fully Connected (FC) layers to learn the time dependencies of affect, and infer the corresponding affective dimension.*

Table 1: *Summary with the number of interactions available from the Cross-cultural Emotion Sub-challenge dataset per culture (German – DE, Hungarian – HU, Chinese – CN) and partition, the duration of the original interactions, and the duration of the compiled segments of the original interactions in which acoustic and visual information corresponds to the same interactant (used in this work). Duration-related information is computed time-wise and displayed with (HH:)MM:SS format.*

Partition	Culture	# Interactions	Duration	
			Original	Used
Train	DE	34	1:33:27	44:03
	HU	34	1:08:41	31:44
	$\sum$	68	2:42:08	1:15:47
Dev.	DE	14	37:52	18:18
	HU	14	28:50	14:34
	$\sum$	28	1:06:42	32:52
Test	DE	16	46:47	21:24
	HU	18	36:18	18:09
	CN	70	3:18:14	1:27:31
	$\sum$	104	4:41:19	2:07:04

information from both. To ensure a fair use of the involved modalities, we exclusively analyse the segments of the interactions corresponding to the timestamps in which the information from both acoustic and visual modalities match; i. e. , the information from the interactant speaking is the same as the one being video recorded. Table 1 summarises the data available and used in this work.

3. Methodology

This section introduces the system implemented (cf. Section 3.1), illustrated in Figure 1, and describes the SSL approach followed in this work (cf. Section 3.2).

3.1. Implemented System

Three main components form our system (cf. Figure 1), which we proceed to describe in the following paragraphs.

Data Preparation. Based on the nature of the CES dataset (cf. Section 2), we first cropped the original videos selecting the timestamps in which the interactant speaking is the same as the one being video recorded. Furthermore, we compensated the delay annotators might have experienced between perceiving and reporting the emotional state of the interactant [24]. Using annotation delay compensation, we shifted the affect-related annotations back in time by 2.4 seconds [25]. The next step is the extraction of audiovisual features from the cropped videos.

The 23 *Low Level Descriptors* (LLDs) of the EGEMAPS feature set [26] are extracted from the audio signals using OPENSMILE [27]. For the visual modality, we opted for extracting the intensities of 17 *Facial Action Units* (FAUs) using OPENFACE [28]. Both acoustic and visual LLDs are extracted at different sampling rates. To overcome this issue, we computed their functionals, as a technique for summarising the information extracted. Specifically, we used sliding windows of 4 seconds length with a hop size of 0.1 seconds to compute the mean and standard deviation of the LLDs extracted in the corresponding time span. The window length selected ensures capturing useful affect-related information [22]. The hop size used contributes to homogenising the sampling rates between the audiovisual functionals and the annotations. The functionals are finally standardised to boost the convergence when training the models.

Affective states are context-related, and, as a consequence, it is beneficial to include contextual information, as past information in the time domain, when modelling affect [29, 30]. This temporal modelling can be achieved using *Recurrent Neural Networks* (RNNs). In this work, we emulated the time annotators need to perceive affect and modelled the current annotation, $y[n]$, with current and previous input features, $\big[x[n], \cdots, x[n-N]\big]$. Nevertheless, the current annotations do not only correlate with the features themselves, but also depend on the previous annotations. Hence, we modelled affective annotations as

$$y[n] = f\left(\begin{bmatrix} x[n] & \cdots & x[n-N] \\ y[n-1] & \cdots & y[n-N-1] \end{bmatrix}\right), \tag{1}$$

where N corresponds to the number of samples needed to capture 2.4 seconds of data, in concordance with our chosen annotation delay compensation factor. This many-to-one approach can be interpreted as a technique for data augmentation.

Neural Network Architecture. Affective annotations are modelled with a *Gated Recurrent Unit Recurrent Neural Network* (GRU-RNN) followed by three stacked *Fully Connected* (FC) layers. The GRU-RNN, with 32 hidden units, aims to capture the time dependencies of the input data sequence, and learns a hidden representation. The purpose of the FC layers, with 32, 16, and 1 neurons, respectively, is to progressively compress the information embedded in the hidden representation learnt with the GRU-RNN. The last FC layer uses a HardTanh activation function, so the inferred annotations belong to the range $[-1, 1]$.

Table 2: *Summary of the Concordance Correlation Coefficients (CCC) obtained by comparing the ground truth and the predicted arousal annotations from acoustic features per culture on both development and test partitions. Specifically, we compared the performance of the models when trained using **M**anual or **A**utomatic annotations as training material. For each scenario, the selection of the interactions used to train the **M** model was performed culture-wise. The highest CCC scores per culture in each scenario are highlighted.*

Culture		DE		HU		DE+HU	
Models		M	A	M	A	M	A
Dev.	DE	**.266**	.059	**.115**	.006	**.219**	.111
	HU	**.019**	.007	.074	**.147**	.075	**.160**
Test	DE	**.102**	.021	**.218**	.027	**.258**	.129
	HU	**.177**	.031	**.200**	.030	**.163**	.040
	CN	**.004**	.003	.007	**.008**	**.007**	.003

The network is trained using the *Concordance Correlation Coefficient* (CCC) as the loss, with Adam as the optimiser. The learning rate of the optimiser was set to $1 \cdot 10^{-4}$. Data from all available interactions was read at once, and we selected one in every five consecutive windows of features as training material. This way, we reduced the oversampling of the training data, and contributed to a better network generalisation. The weights of the network were updated using mini-batches of 1 000 samples. The network was trained during a maximum of 300 epochs, and implemented an early stopping method to stop training when the loss on the validation partition does not improve for 20 consecutive epochs.

As the previous gold standard annotations defined in Equation (1) are not available at inference time, the inferred annotations in previous time steps are used on the prediction of the current annotation. The buffer with previously inferred annotations is initialised with zeros at every new interaction coming to the system, and continuously updated.

Post-processing. The inferred annotations are post-processed using a median filter before the actual assessment of the models. The median filter uses a kernel size of 3 samples to post-process the annotations associated to the audio modality, and a kernel size of 33 samples to post-process the annotations associated to the video modality. These parameters were optimised for the assessment of the baseline model on the development partition.

3.2. Semi-Supervised Learning Approach

Our purpose is to assess the cultural influence on training affect recognition models with SSL annotations. Hence, interactions with SSL annotations need to be included as training material. For a fair comparison between the models, we split the interactions in the train partition into two disjoint subsets, named S_M and S_A. The subset S_M contains half of the original interactions with their corresponding manual annotations, and is used to train a ***M**anual* model. The ***M*** model is then used to automatically annotate the interactions belonging to the subset S_A, which contains the interactions excluded from S_M. Next, we used the interactions belonging to S_A and their corresponding SSL annotations to train an ***A**utomatic* model. Finally, we combined S_M and S_A subsets with their corresponding manual and SSL annotations, respectively, to train a ***M**anual + **A**utomatic* model.

In order to investigate the cultural impact on the performance of SSL, we set two different scenarios. In the first scenario, only German interactions were included in S_M. In the second one, only Hungarian interactions were included in S_M. We extended this analysis with a third scenario, in which S_M contained half of the interactions from both German and Hungarian cultures. This splitting was performed by seeding the pseudo-random number generator and is publicly available[1]. Interactions belonging exclusively to the train partition were used to train the models assessed on the development partition. The models assessed on the test partition used interactions from both train and development partitions as training material. Thus, at this stage, the interactions belonging to the development partition were also split and included in the two disjoint subsets S_M and S_A, and processed as described in the aforementioned procedure.

Table 3: *Summary of the Concordance Correlation Coefficients (CCC) obtained by comparing the ground truth and the predicted arousal annotations from acoustic features per culture on both development and test partitions. The baseline model was trained using the original manual annotations. The remaining models were trained combining manual and automatic annotations (**M+A** model). The interactions used to infer the automatic annotations were selected culture-wise. The highest CCC scores per culture among the hybrid models assessed are highlighted.*

Culture		Baseline	DE	HU	DE+HU
Dev.	DE	.308	**.235**	.010	.224
	HU	.148	.012	.148	**.314**
Test	DE	.142	.099	.036	**.237**
	HU	.165	.203	.052	**.204**
	CN	.006	.005	**.009**	.006

4. Experimental Results

The interactions belonging to the CES dataset had been manually labelled in terms of valence and arousal. Thus, we used these manual annotations to train the baseline models for our experiments. As arousal information is considered to be stronger in the voice, while valence information, in the face [31], we focused our analysis on the automatic recognition of arousal from acoustic features (cf. Section 4.1), and valence from visual features (cf. Section 4.2). The performance of the trained models in the different scenarios outlined in Section 3.2 is assessed by computing the CCC between the inferred and ground truth annotations from all interactions belonging to each specific cultural subset in the development or test partitions.

4.1. Arousal Recognition from Acoustic Information

The results obtained on the automatic recognition of arousal are summarised in Tables 2 and 3. Table 2 compares the performance of the models when using manual or automatic annotations exclusively as training material. Table 3 compares the performance of the baseline model, which uses manual annotations from all the interactions as training material, with the hybrid models, which are trained using both manual and automatic annotations.

The performance analysis of the models trained with manual or automatic annotations (cf. Table 2) indicates the suitability of the manual annotations. When only German interactions were used in S_M, the trained ***M*** model achieved a better performance than the ***A*** model on both development and test partitions. In the second scenario, in which only Hungarian interactions pop-

[1] https://github.com/EIHW/AVEC19CES_CrossCulturalSSL

Table 4: *Summary of the Concordance Correlation Coefficients (CCC) obtained by comparing the ground truth and the predicted valence annotations from visual features per culture on both development and test partitions. Specifically, we compared the performance of the models when trained using* ***M****anual or* ***A****utomatic annotations as training material. For each scenario, the selection of the interactions used to train the* ***M*** *model was performed culture-wise. The highest CCC scores per culture in each scenario are highlighted.*

Culture		DE		HU		DE+HU	
Models		M	A	M	A	M	A
Dev.	DE	.447	**.487**	**.211**	.131	**.360**	.207
	HU	.130	**.161**	**.241**	.168	.182	**.190**
Test	DE	.272	**.357**	**.166**	.117	**.203**	.193
	HU	.100	**.268**	**.085**	.064	**.238**	.078
	CN	.027	**.246**	**.030**	.024	**.043**	.035

ulated S_M, ***M*** and ***A*** models obtained the best performances on the German and Hungarian interactions belonging to the development partition, respectively. On the test partition, the ***M*** model scored the highest CCC on the German and Hungarian interactions, while for the Chinese interactions, the best CCC was obtained with the ***A*** model. In the last scenario, which combined German and Hungarian interactions in S_M, the ***M*** and ***A*** models scored the highest CCC on the German and Hungarian interactions belonging to the development partition, respectively. On the test partition, the ***M*** model obtained the best results in all the cultures assessed. From a cultural perspective, the German model obtained the best performance on the German interactions belonging to the development partition, while the multicultural model scored the highest CCC on the Hungarian ones. On the test partition, the multicultural model achieved the best performance on the German interactions, while the Hungarian model scored the highest CCC on both the Hungarian and Chinese ones.

From the evaluation of the hybrid models (cf. Table 3), we observe that for the three cultures belonging to the test partition, the performance of the best hybrid models surpassed the baseline model. Specifically, hybrid models trained with German and Hungarian interactions in S_M achieved the highest CCC scores on both the German and Hungarian interactions on the test set. On the Chinese culture, the best model was obtained when using Hungarian interactions only in S_M.

4.2. Valence Recognition from Visual Information

The results obtained on the automatic recognition of valence are summarised in Tables 4 and 5. Table 4 compares the performance of the models when using manual or automatic annotations exclusively as training material. Table 5 compares the performance of the baseline model, which uses manual annotations from all the interactions as training material, with the hybrid models, which are trained using both manual and automatic annotations.

The performance analysis of the models trained with manual or automatic annotations (cf. Table 4) shows interesting results. When only German interactions were used in S_M, the ***A*** model obtained the highest CCC scores in all cultures from both development and test partitions. On the other hand, when only Hungarian interactions were used in S_M, the ***M*** model achieved the highest CCC scores in all cultures from both development and test partitions. Finally, when both German and Hungarian interactions populated S_M, the ***M*** and ***A*** models scored the highest CCC on the German and Hungarian interactions belonging to the development partition, respectively. On the test partition, the ***M*** model obtained a better performance than the ***A*** model in all the cultures assessed. From a cultural perspective, the German model obtained the best performance on the German interactions belonging to the development partition, while the Hungarian model scored the highest CCC on the Hungarian ones. On the test partition, the German model scored the highest CCC in all cultural interactions.

Table 5: *Summary of the Concordance Correlation Coefficients (CCC) obtained by comparing the ground truth and the predicted valence annotations from visual features per culture on both development and test partitions. The baseline model was trained using the original manual annotations. The remaining models were trained combining manual and automatic annotations (****M+A*** *model). The interactions used to infer the automatic annotations were selected culture-wise. The highest CCC scores per culture among the hybrid models assessed are highlighted.*

Culture		Baseline	DE	HU	DE+HU
Dev.	DE	.451	**.428**	.183	.310
	HU	.232	.160	.223	**.244**
Test	DE	.233	**.492**	.106	.200
	HU	.153	**.142**	.080	.115
	CN	.039	**.075**	.029	.044

From the evaluation of the hybrid models (cf. Table 5), we observe that for the three cultures belonging to the test partition, the highest CCC scores were obtained with the hybrid model that used German interactions to populate S_M. Specifically, for the German and Chinese interactions, the performance of this model surpassed the baseline model.

5. Conclusions

This work assessed the impact of culture when using SSL on the continuous recognition of affect. Specifically, we focused on the automatic recognition of arousal from the voice, and valence from the face. The results obtained conveyed that the culture of the interactions used for training the models impacted the overall system performance. In most of the cases analysed when comparing ***M*** and ***A*** models, the best performances were obtained when affective models were trained using manual annotations. Nonetheless, the use of SSL annotations alone showed highly competitive results. When analysing the ***M+A*** models, we observed that hybrid solutions, combining manual and automatic annotations, surpassed the baseline, which only used manual annotations, in most of the cases investigated. These results encourage the use of automatic annotations or hybrid solutions to ease the data annotation process in affect-related problems.

Future directions to carry on this work include the cross-modal study of SSL for continuous affect recognition, and the investigation of multi-task networks in this problem in order to exploit the supplementary information embedded in the valence and arousal dimensions simultaneously. Further work can be performed towards a deep understanding of the benefits of using teacher forcing strategies in multimodal paradigms aiming at the continuous recognition of affect.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

7. References

[1] B. Mesquita and N. H. Frijda, “Cultural variations in emotions: A review,” *Psychological Bulletin*, vol. 112, no. 2, pp. 179–204, 1992.

[2] H. A. Elfenbein and N. Ambady, “Universals and Cultural Differences in Recognizing Emotions,” *Current Directions in Psychological Science*, vol. 12, no. 5, pp. 159–164, 2003.

[3] Z. Zhang, N. Cummins, and B. Schuller, “Advanced data exploitation in speech analysis – an overview,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, 2017.

[4] F. Nie, L. Tian, R. Wang, and X. Li, “Multiview Semi-Supervised Learning Model for Image Classification,” *IEEE Transactions on Knowledge and Data Engineering*, 2019, 12 pages.

[5] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, “Improving Landmark Localization With Semi-Supervised Learning,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, 2018, pp. 1546–1555.

[6] V. Kumar, A. Namboodiri, and C. Jawahar, “Semi-supervised annotation of faces in image collection,” *Signal, Image and Video Processing*, vol. 12, pp. 141–149, 2018.

[7] N. F. F. D. Silva, L. F. S. Coletta, and E. R. Hruschka, “A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning,” *ACM Computing Surveys*, vol. 49, no. 1, 2016, 26 pages.

[8] R. Hsiao, T. Ng, F. Grézl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. Schwartz, “Discriminative semi-supervised training for keyword search in low resource languages,” in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*. Olomouc, Czech Republic: IEEE, 2013, pp. 440–445.

[9] A. Kannan, K. Chen, D. Jaunzeikare, and A. Rajkomar, “Semi-supervised learning for information extraction from dialogue,” in *Proceedings of Interspeech*. Hyderabad, India: ISCA, 2018, pp. 2077–2081.

[10] F. Grézl and M. Karafiát, “Combination of Multilingual and Semi-Supervised Training for Under-Resourced Languages,” in *Proceedings of Interspeech*. Singapore: ISCA, 2014, pp. 820–824.

[11] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, “Semi-supervised active learning for sound classification in hybrid learning environments,” *PLOS One*, vol. 11, no. 9, 2016, 23 pages.

[12] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, “Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media,” in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. Sydney, Australia: ACM, 2017, pp. 1191–1198.

[13] A. Mahdhaoui and M. Chetouani, “Emotional Speech Classification Based on Multi View Characterization,” in *Proceedings of the International Conference on Pattern Recognition*. Istanbul, Turkey: IEEE, 2010, pp. 4488–4491.

[14] I. Cohen, N. Sebe, F. G. Cozman, and T. S. Huang, “Semi-Supervised Learning for Facial Expression Recognition,” in *Proceedings of the 5^{th} International Workshop on Multimedia Information Retrieval*. Berkeley, California: ACM, 2003, pp. 17–22.

[15] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, “Learning Bayesian Network Classifiers for Facial Expression Recognition using both Labeled and Unlabeled Data,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. Madison, WI: IEEE, 2003, 7 pages.

[16] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schüller, “Enhanced semi-supervised learning for multimodal emotion recognition,” in *Proccedings of the International Conference on Acoustics, Speech and Signal Processing*. Shanghai, China: IEEE, 2016, pp. 5185–5189.

[17] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, “Cooperative Learning and its Application to Emotion Recognition from Speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.

[18] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, “Leveraging Unlabeled Data for Emotion Recognition With Enhanced Collaborative Semi-Supervised Learning,” *IEEE Access*, vol. 6, pp. 22 196–22 209, 2018.

[19] S. Hantke, A. Abstreiter, N. Cummins, and B. Schuller, “Trustability-based Dynamic Active Learning for Crowdsourced Labelling of Emotional Audio Data,” *IEEE Access*, vol. 6, pp. 42 142–42 155, 2018.

[20] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, “CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms,” in *Proceedings of the 7^{th} International Conference on Affective Computing and Intelligent Interaction*. San Antonio, TX: IEEE, 2017, pp. 340–345.

[21] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[22] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, “AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition,” in *Proceedings of the 9^{th} International Workshop on Audio/Visual Emotion Challenge*. Nice, France: ACM, 2019, pp. 3–12.

[23] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller, K. Star, E. Hajiyev, and M. Pantic, “SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2019.

[24] S. Mariooryad and C. Busso, “Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators,” *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[25] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge,” in *Proceedings of the 6^{th} International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, The Netherlands: ACM, 2016, pp. 3–10.

[26] 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*, vol. 7, no. 2, pp. 190–202, April 2016.

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

[28] T. Baltrušaitis, P. Robinson, and L. Morency, “OpenFace: an Open Source Facial Behavior Analysis Toolkit,” in *Proceedings of the Winter Conference on Applications of Computer Vision*. Lake Placid, NY: IEEE, 2016, 10 pages.

[29] R. W. Levenson, “Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity,” *Social psychophysiology: Theory and clinical applications*, pp. 17–42, 1988.

[30] H. Gunes and M. Pantic, “Automatic, Dimensional and Continuous Emotion Recognition,” *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[31] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey,” in *Proceedings of the 9^{th} International Conference on Automatic Face and Gesture Recognition*. Santa Barbara, CA: IEEE, March 2011, pp. 827–834.