INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Continuous Emotion Recognition in Speech – Do We Need Recurrence?

Maximilian Schmitt[1], *Nicholas Cummins*[1], *Björn Schuller*[1,2]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
[2]GLAM – Group on Language, Audio & Music, Imperial College London, U. K.

maximilian.schmitt@informatik.uni-augsburg.de

Abstract

Emotion recognition in speech is a meaningful task in affective computing and human-computer interaction. As human emotion is a frequently changing state, it is usually represented as a densely sampled time series of emotional dimensions, typically arousal and valence. For this, recurrent neural network (RNN) architectures are employed by default when it comes to modelling the contours with deep learning approaches. However, the amount of temporal context required is questionable, and it has not yet been clarified whether the consideration of long-term dependencies is actually beneficial. In this contribution, we demonstrate that RNNs are not necessary to accomplish the task of time-continuous emotion recognition. Indeed, results gained indicate that deep neural networks incorporating less complex convolutional layers can provide more accurate models. We highlight the pros and cons of recurrent and non-recurrent approaches and evaluate our methods on the public SEWA database, which was used as a benchmark in the 2017 and 2018 editions of the Audio-Visual Emotion Challenge.

Index Terms: affective computing, speech emotion recognition, human-computer interaction, computational paralinguistics, convolutional neural networks

1. Introduction

Emotion recognition in speech (ERS), is a well-studied field in the domain of affective computing [1]. Using the speech signal as a modality has some advantages compared to the visual modality, e. g., that there are no occlusions [2, 3]. Nevertheless, the best performances are usually obtained when following a multi-modal approach fusing the acoustic, linguistic, and visual domains [4]. Whereas early research in ERS pursued the recognition of human emotions in terms of categories (e. g., *happy*, *bored*, etc.) [5], dimensional models have now been established, based on the *circumplex* model introduced by Russell [6]. In these models, emotions are generally expressed in terms of two continuous variables, namely *arousal* and *valence* [7]. While the first describes the level of the physical or mental response of a person, the latter describes whether the emotion is a positive or a negative one. Moreover, ERS is nowadays not commonly performed on discrete chunks of audio, but rather in a 'time-continuous' manner, assigning a level of arousal and valence to each time-stamp of an audio (or audio-visual) data stream [8]. This approach takes into account the rapidly changing nature of emotion as a human state.

From the methodological point of view, time-continuous ERS is usually accomplished by first extracting meaningful acoustic features and then employing a *recurrent neural network* (RNN) architecture, mostly *long short-term memory (LSTM)*-RNNs, to model arousal and valence over time [8, 9, 10]. Instead of building on hand-crafted acoustic feature sets [11], which are motivated by both domain-knowledge and experimental evaluation, also end-to-end learning has become common practice, using a neural network of, for example, two *convolutional layers* (followed by maximum-pooling layers) to handle the feature extraction on the raw time-domain signal, followed by LSTM layers [12]. Convolutional layers can be interpreted as *finite impulse response* filters, where all weights are trained on the data.

Architectures employing LSTM units have the property of modelling the dynamics of long-term time series [13], however, it is questionable if such dependencies are actually required for emotion modelling. Huang et al. found that even with an LSTM model, it is difficult to benefit from long-term patterns in audio-visual emotion recognition, and they trained their model on sub-sequences from the recordings [14]. Furthermore, as emotion is a human state varying quickly over time, it should be possible to recognise it also from a short context window, at least with an acceptable error. This effect has already been demonstrated, using a *support vector machine (SVM)* model based on *bag-of-audio-words* features summarising audio descriptors from a certain block (6 to 8 seconds) as a histogram representation [15, 16]. Nevertheless, recurrent architectures trained on long sequences have the inherent advantage of compensating temporal delays between the recording and the *gold standard* of arousal and valence. This delay originates from the observation that for the labels found in most corpora, human annotators are engaged to rate the emotion in the recordings based on their perceptual observations while listening and/or watching. The human decision-making process inherently takes a certain amount of time, typically 2 to 4 seconds [15, 17, 18]. The resulting delay of the annotated contours must be compensated for when a non-recurrent approach or a static regressor, such as SVM, is employed.

In this contribution, we demonstrate that a recurrent architecture is actually not beneficial for time-continuous ERS and that a deep neural network architecture consisting of only convolutional layers provides superior results. This architecture has the advantage of having less trainable parameters and, therefore, a faster training process. Both the emotion predictions in an intra-cultural and in a cross-cultural setting are improved, compared to the recurrent model. Moreover, we can show that the training process is quite robust against delays between the audio signal and the gold standard. To the knowledge of the authors, this is the first work on *time-continuous* ERS using a deep learning architecture with a fully convolutional neural network (CNN). In previous works, such as by Zheng et al., *categorical* emotion in audio *chunks* is classified with a CNN, based on a whitened spectrogram representation [19].

In the following section, the corpus used throughout the experiments and related work on it are introduced. In Section 3, the features and models we compare are motivated and

http://dx.doi.org/10.21437/Interspeech.2019-2710

Table 1: *Key statistics of the AVEC 2018 CES data [21]*

Culture	Partition	#Subjects	Length [min:s]
German	Training	34	93:12
German	Development	14	37:46
German	Test	16	46:38
Hungarian	Test	66	133:12
SUM		130	310:48

described. Next, in Section 4, experiments and results are presented and discussed in Section 5. Finally, we conclude and give an outlook on future research in Section 6.

2. Corpus

The SEWA corpus includes audio-visual recordings of subjects of different cultures watching and discussing commercials through an online platform [20]. In this study, we use the *German* and *Hungarian* video chats from the SEWA corpus, where a pair of subjects discusses the last commercial seen beforehand in an up to 3-minutes-long video chat. This subset has already been used as a benchmark for the 2017 and 2018 editions of the Audio-Visual Emotion Challenge (AVEC) [4, 21]. In the AVEC 2018 Cross-cultural Emotion Sub-Challenge (CES), participants were invited to create a model for *arousal*, *valence*, and *liking (sentiment)* on provided Training and Development sets from the German culture and submit their predictions on a German Test set and the whole Hungarian set, for which no gold standard annotations were provided. For all experiments described in this contribution, exactly the same setting is used as for the AVEC 2018 CES challenge. The statistics of the dataset are given in Table 1. The data was annotated by 6 (German) and 5 (Hungarian) annotators of the respective culture and the single ratings were fused to a unique *gold standard* for each dimension using a variant of the *evaluator weighted estimator*, with an output frequency of 10 Hz [4, 21]. In AVEC 2018 CES, a model based on different (hand-crafted and deep) acoustic and facial features and a 2-layer LSTM is proposed as a baseline.

Time-dependent modelling was widely proposed in the contributions to AVEC. In the approach by Wataraka Gamage et al., emotional dimensions are modelled as the outputs of time-invariant filter arrays, each filter representing a ‘salient event’ [22]. Huang et al. employ a fusion of different (hand-crafted and deep) feature sets and an LSTM-RNN [14] and investigate on data augmentation by cutting and overlapping the long sequences. The AVEC 2018 CES winners, Zhao et al. [23] use features from a pretrained deep model (VGGish [24]) for audio and an LSTM-RNN. Besides, approaches using only hand-crafted and no deep audio features have achieved a good performance in the past [25, 26]. Han et al. showed on the SEWA corpus that multi-task learning, learning arousal and valence contours together and using also the uncertainty between annotators as additional targets, improved the models [27]. A numerical overview over the results obtained on the SEWA datasets in the AVEC 2017 and 2018 settings is given in Section 4. For the study presented in this contribution, we focus on the audio modality and on the prediction of the emotional dimensions *arousal* and *valence* as the recognition of *liking* typically requires the inclusion of explicit linguistic cues [21].

3. Models

The investigated models are based on hand-crafted acoustic features. We compare two deep neural network (DNN) models: a model consisting of LSTM layers and a model consisting of convolutional layers only. The features and models are explained in the following.

3.1. Acoustic features

It has previously been demonstrated that features representing the *mean* and the *standard deviation* of the acoustic low-level descriptors (LLDs) defined in eGeMAPS[1] [11] outperform both the functionals defined in eGeMAPS and *bag-of-audio-words representations* when used as an input for an LSTM-based emotion recognition model [25]. Furthermore, it was also shown that for those supra-segmental features, a small window size of 100 ms, over which the *mean* and *standard deviation* are computed is well suited for an LSTM-backend. We use exactly the same input as a baseline for the proposed approach. The 23 eGeMAPS LLDs are extracted for all audio files using the toolkit OPENSMILE [28]. They include an expert-defined set of acoustic descriptors relevant for affect, such as *Mel-frequency cepstral coefficients*, *pitch*, *formant frequencies*, or *jitter & shimmer* [11]. The mean and standard deviation for each LLD are computed over the mentioned windows with a step size of 100 ms to match with the gold standard annotations. As both speakers of each session are present in a single audio recording [21], we exploit the information on turn timings as a single additional feature denoting speaker presence (1.0 or 0.0) as in the work by Huang et al. (“mark method”) [14, 25]. Therefore, in total, we use a 47-dimensional feature vector for each 100 ms step.

3.2. DNN architectures

3.2.1. LSTM model

The LSTM model is similar to the one used in our previous work [25], achieving a performance comparable to the top performing models of the AVEC challenges when considering only the acoustic domain [29]. The model consists of 4 LSTM layers with the default *tanh* activation after each layer. As the temporal delay of the gold standard is optimised during the experiments, i. e., the target contour is shifted back in time for a specified interval prior to training, the choice of a *uni-directional* LSTM architecture would not be fair, as it takes into account only past context, which is shrinking with increasing delay compensation. This effect is why, in contrast to the previous work [25], all LSTM layers are *bi-directional*, i. e., they capture both past and future context of the sequence. Thus, at each time step, any information from the whole sequence can have an effect on each prediction. Neither dropout nor batch normalisation improved the performance of the model. A single output neuron with a *linear* activation is used as our initial experiments revealed that the multi-target learning of arousal and valence was not beneficial in this setting.

3.2.2. CNN model

The CNN model consists of 4 convolutional layers with a *ReLU* activation and a single output neuron with a *linear* activation. In initial experiments, a 4-layer CNN showed a performance superior to a 2-layer architecture and a slighly better performance than a 3-layer architecture. Each layer has an increasing filter length spanning 5, 20, 30, and 50 time steps, respectively. Considering all layers, for each time step, the last layer receives a context of almost 8 s from the audio signal, given that the first and last outputs of the 1^{st}, 2^{nd}, and 3^{rd} convolutional layers increase the receptive field of

[1] extended Geneva Minimalistic Acoustic Parameter Set

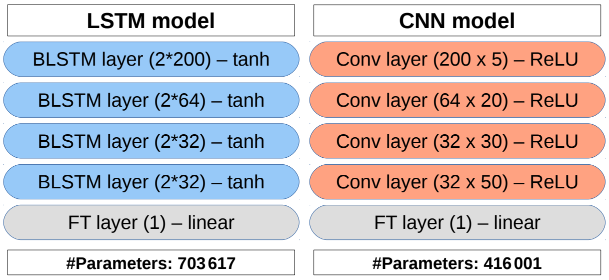

LSTM model
BLSTM layer (2*200) – tanh
BLSTM layer (2*64) – tanh
BLSTM layer (2*32) – tanh
BLSTM layer (2*32) – tanh
FT layer (1) – linear
#Parameters: 703 617
CNN model
Conv layer (200 x 5) – ReLU
Conv layer (64 x 20) – ReLU
Conv layer (32 x 30) – ReLU
Conv layer (32 x 50) – ReLU
FT layer (1) – linear
#Parameters: 416 001

Figure 1: *Overview of the LSTM and CNN models. BLSTM: Bi-directional Long Short-Term Memory, Conv: Convolutional, FT: Fully-connected Time-distributed. Numbers in brackets denote the number of units/filters, the second number in the convolutional layers denotes the filter length (temporal context).*

the last layer by the sum of their halved filter lengths. This amount of temporal context has proven to be suitable already in previous ERS research with static classifiers such as SVM [15]. Maximum pooling is not applied anywhere as the input features and the target annotations have the same step size of 100 ms. As for the LSTM model, neither dropout nor batch normalisation, and only a single output were employed.

An overview over both the LSTM and the CNN model with the number of units in each layer is given in Figure 1. Even though it is difficult to compare the number of LSTM units and CNN filters as different numbers of model parameters are connected to those, we choose the same number of LSTM units and CNN filters in each layer. It must be pointed out that the number of units applied to both the forward-directed and the backward-directed layer, so that the final number of units is double. In initial experiments, we observed that only using half of the number of units lead to worse results for the LSTM model. The total number of parameters is also given in Figure 1.

4. Experiments and results

The DNNs are implemented in the KERAS (v2.2.4) framework with TENSORFLOW backend (v1.12.0). Training and evaluation are done on an NVIDIA GEFORCE GTX 745 consumer graphics card (CUDA version 9.0.176, cuDNN version 7.3.1). For all LSTM layers, the very fast CuDNNLSTM implementation from KERAS is used. Training is done on the full batch of 34 sequences and ruled by the *ADAgrad* optimiser, with an initial learning rate of 0.001, which is close to optimum based on previous and further experiments [25]. The number of epochs is fixed to 500, where no improvement is found anymore for both models, and the network weights from the epoch with the lowest loss on the Development set are restored to predict on the Test sets of both cultures. No post-processing is applied to the output predictions.

The *concordance correlation coefficient* (CCC) is used as both the objective function for training and as the evaluation metric, where a CCC of 1.0 means perfect prediction and a CCC of 0.0 represents chance level. To be consistent with the protocol applied in the AVEC 2018 CES [21], the CCC is computed on the concatenated sequences for each evaluated partition. As mentioned before, different models are trained for arousal and valence as multi-target learning did not provide any improvement in the given setting.

In the first round of experiments, the delay compensation is optimised. For this, the arousal/valence gold standards were shifted temporally towards the front for a certain number of seconds (0.0, 0.4, ..., 6.0) for training and the predictions were shifted back in time for the same interval. The results on the Development set are displayed in Figure 2 for both models and each emotional dimension. The evaluation shows that, surprisingly, the delay compensation does not have a considerable influence on the results for the CNN model, where the CCC is quite stable over a broader range, but rather for the LSTM model. Only for a window size larger than 5.0 s, the performance drops rapidly. This result supports the findings by Huang et al. [14] and shows that LSTMs might not learn any long-term patterns for the given task.

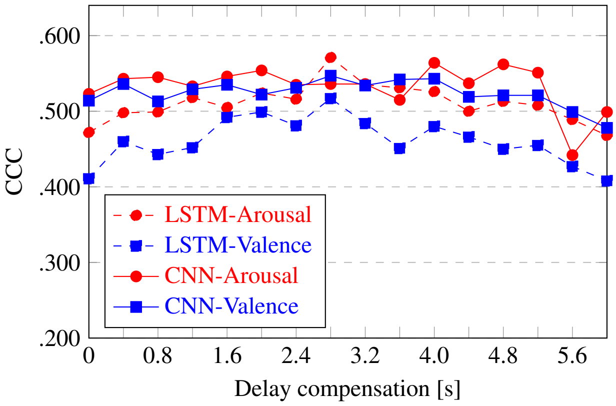


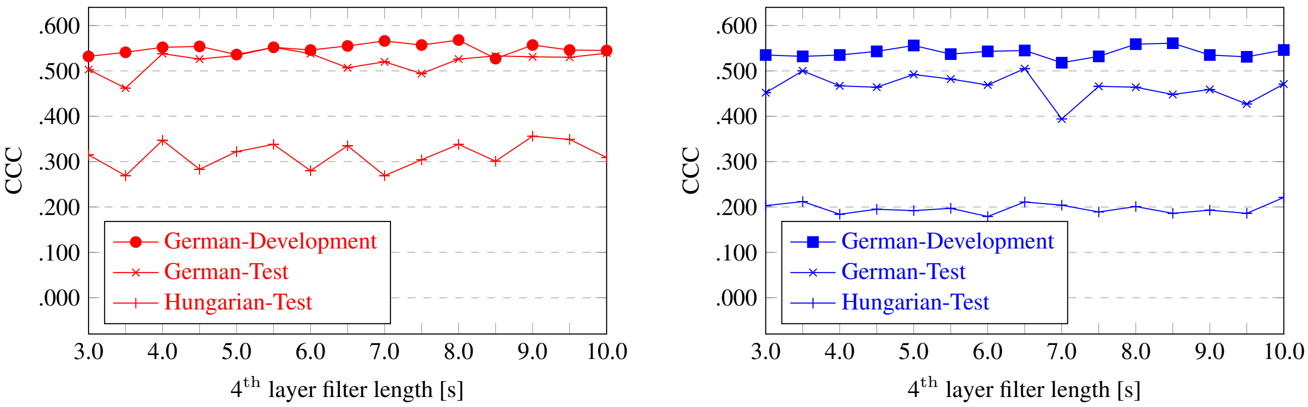
Figure 2: *Optimisation of the delay compensation on the AVEC 2018 CES Development set for both models and arousal/valence.*

The second round of experiments investigated, how the performance of the CNN was affected by varying the filter lengths. To reduce the complexity of the experiments, we only modify the length of the last convolutional layer. As pointed out previously, due to the overlap of the previous layers, the captured context is always a bit larger than the actual filter length. We use the optimised delays for arousal (4.0 s) and valence (2.8 s), respectively. Results from these experiments are shown in Figures 3a (arousal) and 3b (valence) for all partitions. It is evident that also the filter length of the final layer does not have a substantial influence on the final results. Indeed, optimisation of the last layer's filter length provides only a low improvement on the Development set and no improvement on the Test sets, in comparison with the length presumed initially.

Table 2 gives an overview of the results obtained with the investigated models and compares them to those gained with different other approaches from the literature. All Test set results displayed were obtained with the very same model that performed best on the Development set. Some of these approaches were using further modalities, i. e., video or linguistics, based on the ground truth transcriptions of the speech.

5. Discussion

The presented results demonstrate that the proposed CNN architecture outperforms the LSTM architecture in all tasks, except for arousal on the Development and the Hungarian Test set, where matching results are achieved. While it could be argued that the hyperparameters of the LSTM model allow for more tuning, also the multi-modal baselines of AVEC and all tuned systems based on only the audio modality are outperformed on the German Test set for arousal [22, 25, 27]. For valence, our system is superior compared to three other models, considering the intra-cultural evaluation. For the cross-cultural evaluation,



(a) *Arousal*

(b) *Valence*

Figure 3: *Optimisation of the filter length of the* 4^{th} *layer of the CNN model.*

Table 2: *Comparison of results on the AVEC 2018 CES data. All results are given in terms of CCC. Missing figures (–) were not published. GER: German, HUN: Hungarian*

Model	Ref.	Modalities	Arousal			Valence		
			GER Devel.	GER Test	HUN Test	GER Devel.	GER Test	HUN Test
AVEC 2017 Affect Baseline	[4]	optimised	.373	.375	–	.507	.466	–
AVEC 2017 Affect Winners (Chen et al.)	[29]	audio+video+ling.	.823	.675	–	.796	.756	–
AVEC 2018 CES Baseline	[21]	audio+video	.486	.524	.436	.549	.577	.405
AVEC 2018 CES Winners (Zhao et al.)	[23]	audio+video+ling.	.820	.704	.562	.795	.783	.438
		audio	.604	–	–	.511	–	–
Wataraka Gamage et al.	[22]	audio+ling.	.440	.444	.310	.543	.537	.241
Huang et al.	[14]	audio+video+ling.	.699	.599	.456	.756	.721	.403
Han et al.	[27]	audio+video	.559	.450	–	.575	.515	–
		audio	.356	.275	–	.396	.292	–
eGeMAPS-functionals + 4-layer LSTM	[25]	audio	.586	.499	–	.516	.489	–
Proposed LSTM Model	–	audio	.571	.470	.337	.517	.410	.152
Proposed CNN Model (4th layer width 5.0 s)	–	audio	.564	.536	.339	.547	.479	.192
Proposed CNN Model (optimised 4th layer)	–	audio	.568	.526	.338	.561	.448	.186

only the model for arousal by Wataraka Gamage et al. is surpassed, which is based on audio and linguistics, but neither the valence model nor the multi-modal models are surpassed. However, previous work has shown that the linguistic modality provides more meaningful cues for the recognition of valence than the acoustics [4].

The proposed CNN model, of course, has some pros and cons. One advantage of CNNs is that they can be trained very efficiently using state-of-the-art implementations. Training of one epoch on the system described previously took approximately 10 ms (milliseconds), whereas it took around 57 ms for the LSTM model. Nevertheless, this figure also highly depends on the sequence length and as shown in other works, sequences can be truncated easily to speed-up training. Moreover, we found the convergence to be faster with the LSTM model. Decoding is also faster; it took only 3 ms (CNN) and 16 ms (LSTM) for the whole German test set on the described computer system. Thus, the only factor that limits the usage of the models in real-time applications is the amount of required context. As seen in Figure 3, the context for the CNN can be limited to approximately 4.0 s to 5.0 s, resulting in a delay of only 2.5 s, given that half of the context is in the past. In this regard, one aspect that still needs to be investigated, is how a *causal* CNN architecture, which does not take into account future context, affects the performance. The same goes for LSTM, where the performance of a uni-directional LSTM, considering only the past, should also be evaluated further. Finally, we have shown that the CNN model is quite robust against delays of the gold standard, even more robust than LSTM models, which are supposed to inherently account for shifts between input and output.

6. Conclusion and outlook

We have shown that time-continuous emotion recognition in speech does not require a recurrent deep learning architecture, such as LSTM, to be competitive and that a fully convolutional network achieves superior performances. Furthermore, it was proven that neither delays of the gold standard nor the length of the context considered by the CNN requires much optimisation, proving the potential of this approach for real-time systems.

Future research within the framework of this work will investigate end-to-end CNN-only models for emotion recognition and novel strategies to exploit linguistic information.

7. Acknowledgements

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

8. References

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

[2] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[3] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 25:1–25:49, 2018.

[4] 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 *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, CA, USA: ACM, 2017, pp. 3–9.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th European Conference on Speech Communication and Technology*. Lisbon, Portugal: ISCA, 2005.

[6] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[7] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience," *Psychological Bulletin*, vol. 139, no. 4, p. 917, 2013.

[8] 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*, vol. 66, pp. 22–30, 2015.

[9] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Interspeech*. Brisbane, Australia: ISCA, 2008, pp. 597–600.

[10] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. International Joint Conference on Artificial Intelligence*, 2016, pp. 2196–2202.

[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*, vol. 7, no. 2, pp. 190–202, 2016.

[12] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. Shanghai, P. R. China: IEEE, 2016, pp. 5200–5204.

[13] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium: Presses Universitaires de Louvain, 2015, pp. 89–94.

[14] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Multimodal continuous emotion recognition with data augmentation using recurrent neural networks," in *Proc. 2018 on Audio/Visual Emotion Challenge and Workshop*. Seoul, Republic of Korea: ACM, 2018, pp. 57–64.

[15] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. Interspeech*. San Francisco, CA, USA: ISCA, 2016, pp. 495–499.

[16] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. Schuller, "Bags in bag: Generating context-aware bags for tracking emotions from speech," in *Proc. Interspeech*. Hyderabad, India: ISCA, 2018, pp. 3082–3086.

[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 *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*. Brisbane, Australia: ACM, 2015, pp. 41–48.

[18] 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.

[19] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, AAAC. Xian, P. R. China: IEEE, 2015, pp. 827–831.

[20] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," https://arxiv.org/abs/1901.02839, 2019, 17 pages.

[21] 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 *Proc. 2018 on Audio/Visual Emotion Challenge and Workshop*. Seoul, Republic of Korea: ACM, 2018, pp. 3–13.

[22] K. Wataraka Gamage, T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "Speech-based continuous emotion prediction by learning perception responses related to salient events: A study based on vocal affect bursts and cross-cultural affect in AVEC 2018," in *Proc. 2018 on Audio/Visual Emotion Challenge and Workshop*. Seoul, Republic of Korea: ACM, 2018, pp. 47–55.

[23] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proc. 2018 on Audio/Visual Emotion Challenge and Workshop*. Seoul, Republic of Korea: ACM, 2018, pp. 65–72.

[24] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. Calgary, Canada: IEEE, 2017, pp. 131–135.

[25] M. Schmitt and B. Schuller, "Deep recurrent neural networks for emotion recognition in speech," in *Proc. DAGA*, vol. 44. Munich, Germany: DEGA, 2018, pp. 1537–1540.

[26] V. Pandit, N. Cummins, M. Schmitt, S. Hantke, F. Graf, L. Paletta, and B. Schuller, "Tracking authentic and in-the-wild emotions using speech," in *Proc. 1st Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. Beijing, P. R. China: IEEE, 2018, pp. 1–6.

[27] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. 25th ACM International Conference on Multimedia*. Mountain View, CA, USA: ACM, 2017, pp. 890–897.

[28] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM International Conference on Multimedia*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[29] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, CA, USA: ACM, 2017, pp. 19–26.