Multi-corpus Experiment on Continuous Speech Emotion Recognition: Convolution or Recurrence?

Manon Macary, Martin Lebourdais, Marie Tahon, Yannick Estève, Anthony Rousseau

Multi-corpus experiment on continuous speech emotion recognition: convolution or recurrence?

Manon Macary[1,2], Martin Lebourdais[1], Marie Tahon[1], Yannick Estève[2], and Anthony Rousseau[2]

[1] LIUM, Le Mans {martin.lebourdais,marie.tahon}@univ-lemans.fr
[2] Allo-Média, Paris {m.macary,a.rousseau}@allo-media.fr
[3] LIA, Avignon yannick.esteve@univ-avignon.fr

Abstract. Extraction of semantic information from real-life speech, such as emotions, is a challenging task that has grown in popularity over the last few years. Recently, emotion processing in speech moved from discrete emotional categories to continuous affective dimensions. This trend helps in the design of systems that predict the dynamic evolution of affect in speech. However, no standard annotation guidelines exist for these dimensions thus making cross-corpus studies hard to achieve. Deep neural networks are nowadays predominant in the task of emotion recognition. Almost all systems use recurrent architectures, but convolutional networks were recently reassessed as they are faster to train and have less parameters than recurrent ones. This paper aims at investigating pros and cons of the aforementioned architectures using cross-corpus experiments to highlight the issue of corpus variability. We also explore the best suitable acoustic representation for continuous emotion, together with loss functions. We concluded that recurrent networks are robust to corpus variability and we confirm the power of cepstral features for continuous Speech Emotion Recognition(SER), especially for satisfaction prediction. A final post-treatment applied on prediction brings very nice result ($ccc = 0.719$) on AlloSat and achieves new state of the art.

Keywords: Continuous Speech Emotion Recognition · Deep Neural Networks · Acoustic features

1 Introduction

Semantic information extraction from real-life human-human conversations, such as concepts, emotions or intents, is a major topic for speech processing researchers and companies, especially in call-center related activities. In this context, Speech Emotion Recognition (SER) remains unsolved while being a research field of interest for many years. Predicting naturalistic emotions is a challenging task, especially if the scope is to capture subtle emotions characterizing real-life behaviors.

In the discrete theory [3], emotion categories are usually defined on a word, a segment or a conversation level. However, this approach does not permit to

extract the evolution of affect along a conversation. In the continuous theory, the complex nature of emotion in speech is described with continuous dimensions, notably arousal and valence [15], but also dominance, intention, conductive/obstructive axis [16]. As of now, there are only a few available realistic SER corpora, as they need to respect strong ethical and legal issues and collecting emotional speech demands tremendous efforts. Among these, very few has been annotated continuously according to emotional dimensions because it is a difficult and expensive task. Among the most popular, we can cite SEMAINE [11] (English interactions with virtual agent), RECOLA [12] (on-line conversations). The cross-cultural emotion database (SEWA) [8] (human-human conversations) was presented for the 2018 Audio/Visual Emotion Challenge [13] which aimed to retrieve arousal, valence and liking dimensions across different cultures. Recently, AlloSat [10] (French call-centers telephone conversations) was annotated along the satisfaction dimension. The present study aims at continuously predicting affect in naturalistic speech conversations: this explains why we focus on the SEWA and AlloSat databases.

In SER, speech signals are traditionally represented with acoustic vectors which represent the entire emotional segment. Finding the better acoustic feature set is still an ongoing active sub-field of research in the domain [7]. Most of existing sets intend to describe prosody in the signal, with low level descriptors capturing intensity, intonation, rhythm or voice quality. These features have the advantage of being easily interpretable, however their extraction in degraded signals are error-prone. The HUMAINE association also took an inventory of acoustic features in the CEICES initiative [1] which conducts to a set of a hundred of descriptors selected over several corpora with various techniques [5]. Another option is to extract spectral features; mel frequency cepstral coefficients (MFCCs) are clearly the most often used as they are robust to noisy signals. Previous studies have shown that some implementations of these MFCCs perform better than others for SER [18].

To perform SER tasks, deep neural networks are more and more used, especially for continuous emotion recognition. Recurrent neural networks (RNNs), especially long short term memory (LSTM)-based architectures, are particularly convenient thanks to their memory properties, enabling to retrieve the evolution of long-term time series [6] such as emotions or intents. Convolutional neural networks (CNNs) are widely used in image recognition as they are designed to exploit spatially local correlations. This type of network has not been used in continuous SER to the best of the author's knowledge, until [17] in which the authors conclude to the advantage of CNNs over RNNs to continuously predict arousal and valence in SEWA. As CNNs are faster to train and have less parameters to optimize, they could be interesting to use for continuous SER tasks instead of RNNs.

The main difficulty encountered in SER is due to the task itself: emotions are complex, subtle and subjective thus making the reproduction of the results and standardization very hard. Cross-corpus approaches imply to compare automatic predictions given by a single system on different corpora. This task has already

been investigated to retrieve emotion categories [2, 4, 18], but as far as we know, not to retrieve continuous affective dimensions. The standardization problem is addressed by AVEC challenges [13, 14] which aim at defining standard models and evaluations metrics for continuous SER tasks.

A first goal is to estimate in what extent the experiments realised by [17] are consistent on other data. To do so, the two studied datasets SEWA and AlloSat are introduced in section 2. Section 3 compares system architectures (CNN or RNN) while section 4 explores the impact of input acoustic features on performances. In section 5 we also discuss on results and on the introduction of post-treatment applied to the neural network output.

2 Emotional data and acoustic features

2.1 AlloSat

AlloSat [10] is composed of 303 call-centers telephone real-life conversations. Speakers are French native adult callers (i.e. customer) asking for information such as contract information, complaints on multiple domain company (energy, insurance, etc.). Each signal contains only the caller's speech as the part of the receiver (i.e agent) has been discarded from the corpus for ethical and commercial reasons. A continuous annotation among the satisfaction dimension was realized by 3 annotators on all conversations with a time step of 250 ms. The unique gold reference is the mean of each annotator's values. The corpus is divided into train, development and test sets as shown in Table 1 totalling 37h17' of conversations. For each conversation the speaker is different, ensuring a speaker independent partition.

Table 1: Number of mono speaker conversations (and duration) in AlloSat and the two configurations of SEWA (German only or German and Hungarian), in train, development and test sets.

Corpus	Language	**Train**	**Dev**	**Test**
AlloSat	French	201 (25h26')	42 (5h55')	60 (5h58')
SEWA	German	34 (1h33')	14 (37')	82 (3h)
SEWA	Ger + Hun	68 (2h41')	28 (1h05")	104 (4h40')

2.2 SEWA

The cross-cultural Emotion Database (SEWA) [8] consists of 48 audiovisual recordings of elicited reactions between unique pairs of subjects. Pairs are discussing for less than 3 minutes about an advert seen beforehand. The database is now a reference in the community, as it has been used in the two last Audio/Visual Emotion Challenges (AVEC). In this study, only a subset of the database, containing German and Hungarian records, is investigated according to the guidelines of AVEC 2018 and 2019 workshop [13, 14]. A continuous annotation among three dimensions (arousal, valence and liking) was made by 6

annotators and a unique gold reference has been computed, for every 100 ms. The additional liking axis describes how much the subjects liked the commercial. The corpus is divided into train, development and test sets as shown in Table 1. Test gold references are not distributed. Predictions have to be sent to AVEC organizers to get the final performances.

2.3 Acoustic features

This paper mainly tries to reproduce the experiments from AVEC challenges and Schmitt et al.[17] with different data, to analyze the robustness to corpus variability. To do so, the acoustic feature sets used in these challenges are used as well as an additional feature set. In the end, either hand crafted expert features or Mel Frequency Cepstral Coefficients (MFCC) are used in the following experiments. First, low-level descriptors (LLD) are extracted directly from the speech signal, each 10 ms. These LLDs are then summarized over a fixed time window of 100 ms for SEWA and 250 ms for AlloSat.

- **eGeMAPS-88** contains 88 features from the extended Geneva Minimalistic Acoustic Parameter Set [5]. This set consists of LLDs capturing spectral, cepstral, prosodic and voice quality information from the speech signal, which are then summarised over the time window with a set of statistical measures. This feature set has been extensively used for SER, especially thanks to AVEC challenges. This feature set is extracted with the toolkit OpenSmile[1].
- **eGeMAPS-47** is a subset of GeMAPS which includes 23 LLDs. Mean and standard deviation of these 23 LLDs are computed over the time window. This feature set is extracted with the toolkit OpenSmile. An additional binary feature denoting speaker presence extracted from speech turns, is also included.
- **Mfcc-Os** consists of MFCC1-13, and their first and second derivatives, also extracted with the toolkit OpenSmile. Mean and standard deviation are computed on these 39 features over the time window. In total, we use a 78-dimensional feature vector.
- **Mfcc-lib** is an alternative implementation of cepstral coefficients from librosa[2] that is used in many speech processing experiments. In this set, 24 MFCCs are extracted each 10 ms on a 30 ms window and summarized with mean and standard deviation over the time window. In total, we use a 48-dimensional feature vector.

In SER, acoustic representation of emotion usually tends to capture prosody. That is the reason why expert features are more often used than single MFCC. However, in the context of telephone conversations, the audio signal is severely degraded and expert features can not avoid estimation errors. Therefore MFCC features can be more robust and reliable than expert features and thus gives us better performance in this context. Experiments are conducted in Section 4 to analyze the impact of the features set chosen.

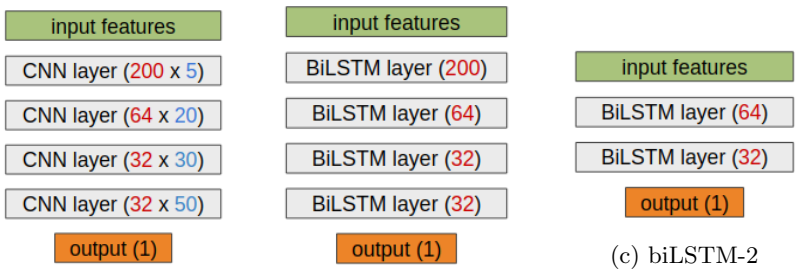
[1] http://audeering.com/technology/opensmile/
[2] https://librosa.github.io/librosa/

3 Convolutional or recurrent models ?

To estimate in what extent the experiments published in [17, 13, 14] are consistent on different corpus, we decided to reproduce them on SEWA and compare them to those obtained on AlloSat.

3.1 Network architectures

In Schmitt et al. [17], CNN and RNN architectures are investigated. The CNN is composed of 4 convolutional layers with a ReLU activation. The RNN is composed of 4 bidirectional Long Short Term Memory (biLSTM-4) layers of respectively 200, 64, 32 and 32 units, and a tanh activation. A single output neuron is used to predict the regression samples each 250 ms for AlloSat, respectively 100 ms for SEWA. In addition, a second RNN with 2 bidirectional LSTM (biLSTM-2) layers of respectively 64 and 32 units, with a tanh activation, proposed in AVEC 2018 and 2019 challenge [13, 14], is experimented. In the end 3 different networks are tested as shown in Fig. 1.



(a) CNN (b) biLSTM-4 (c) biLSTM-2

Fig. 1: Description of the models used: units are in red and filter span in blue.

3.2 Protocol

The networks, summarized in Fig. 1, are implemented with the Keras framework[3] using the Tensorflow backend[4]. The learning rate has been empirically set to 0.001, the number of epochs is fixed to 500 and ADAgrad optimiser is used. The Concordance Correlation Coefficient (CCC) [9] was established as a standard metric in the two last AVEC challenges and is generally used to compute the network loss function. As the random initialization of the system can alter the prediction, each network is trained five times with different seeds and all following results represent the average of these 5 systems and the best score. Table 2 reports the reproduction results in terms of CCC, obtained with the following configurations:

[3] https://keras.io
[4] https://www.tensorflow.org/

- AVEC 2018 (green): network: biLSTM-2; features: Mfcc-Os, eGeMAPS-88; train/dev: SEWA (Ger), AlloSat
- AVEC 2019 (purple): network: biLSTM-2; features: Mfcc-Os, eGeMAPS-88; train/dev: SEWA (Ger+Hun), AlloSat
- Schmitt et al. (pink): networks: CNN, biLSTM-4; features: eGeMAPS-47; train/dev: SEWA (Ger), AlloSat

3.3 Cross-corpus experiments: CNN vs. RNN

Table 2: Comparison of averaged CCC scores of AlloSat and SEWA development sets on the 4 dimensions: satisfaction, arousal, valence and liking with CCC as loss function. Reports of the results from [13, 14, 17] are also included. *Train and prediction has been made on the concatenation of German and Hungarian SEWA conversations.

Models	**Features**	**AlloSat**	**SEWA**		
		satisfaction	arousal	valence	liking
Our systems					
CNN	eGeMAPS-47	.178 (.458)	**.528** (.541)	**.515** (.527)	**.304** (.321)
biLSTM-4	eGeMAPS-47	.437 (.458)	.487 (.527)	.428 (.468)	.258 (.346)
biLSTM-2	eGeMAPS-88	.480 (.564)	.280 (.357)	.174 (.212)	.095 (.171)
biLSTM-2	Mfcc-Os	.364 (.439)	.395 (.438)	.325 (.373)	.158 (.208)
biLSTM-2	eGeMAPS-88*	.480 (.564)	.244 (.273)	.118 (.155)	.082 (.132)
biLSTM-2	Mfcc-Os*	.364 (.439)	.325 (.326)	.186 (.192)	.125 (.126)
biLSTM-4	eGeMAPS-88	**.564** (.634)	.316 (.429)	.237 (.309)	.119 (.188)
Schmitt et al. : Train and Dev on German conversations					
CNN	eGeMAPS-47		.571	.517	
biLSTM-4	eGeMAPS-47		.568	.561	
AVEC 2018 : Train and Dev on German conversations					
biLSTM-2	eGeMAPS-88		.124	.112	.001
biLSTM-2	Mfcc-Os		.253	.217	.136
AVEC 2019 : Train and Dev on German and Hungarian conversations					
biLSTM-2	eGeMAPS-88*		.371	.286	.159
biLSTM-2	Mfcc-Os*		.326	.187	.144

According to Table 2, our models perform in average slightly better than AVEC 2018 baseline on all SEWA dimensions. Although our performances are in average almost below AVEC 2019 baseline, our best models trained with Mfcc-Os features are in the range of the baseline results. This underlines the importance of initialization and seed choice. The comparison with Schmitt et al. shows that we did not managed to reproduce the published results on both arousal and valence dimensions except on the prediction of valence with CNN. We also report the results obtained on liking which are better than the ones obtained with AVEC systems.

Comparison of average and best performances allows us to find the systems which have low dependency with weight initialization, which are biLSTM-2 trained with Mfcc-Os on SEWA Ger+Hun (purple) and CNN trained with

eGeMAPS-47 on SEW Ger (pink). Generally, the best performances seem to be more difficult to reproduce on liking than arousal or valence.

The performances on satisfaction prediction are comparable to those obtained on other dimensions when using biLSTM. However, the results with CNN completely differ. More precisely, with 5 different seeds, 3 models diverged, while all biLSTM converged to satisfactory results. It shows that satisfaction prediction performs better with RNNs than CNNs. Therefore, continuous SER does require a biLSTM architecture to be competitive on multiple dimensions and resolve the corpus variability issue.

We noticed that satisfaction (resp. liking) varies very slowly (resp. slowly) in time in comparison to arousal and valence. This can be due to the annotation protocol (mouse vs. joystick) and the affective content. To investigate if the dynamics in annotation was responsible for poor results on satisfaction, we run additional experiments with smoothed references. The results (not reported here) show that smoothing the reference up to 1 s helps in increasing performances on AlloSat (average $ccc = 0.185$, best $ccc = 0.475$) with the systems which were already converging. To conclude, CNN can not be used on every kind of speech data as their convergence is not straightforward, probably because of their dependency on filter initialization.

Focusing on satisfaction prediction , eGeMAPS-88 ($ccc = 0.480$) appears to perform better than MFcc-Os ($ccc = 0.364$) with 2 biLSTM layers in the network. However, eGeMAPS-47 with 4 biLSTM layers also reaches good results ($ccc = 0.437$). To conclude on the best number of layers, we run a final experiment with 4 biLSTM layers and eGeMAPS-88 ($ccc = 0.564$) which achieves our best result.

This first experiment concludes that biLSTM-4 is the best architecture regarding variability robustness over different emotion corpora. However, the structure of the network is not the only component implicated in a SER module. The representation of the speech data in input and the loss function are also crucial. The following section explores acoustic features and loss functions, using biLSTM-4 models.

4 Impact of input features and loss function

In this section, we study the impact of the acoustic representation of speech in input together with loss functions used during the training phase.

4.1 Loss functions

Traditionally in continuous emotion prediction, the loss function is computed with the CCC as this metric was established as a standard metric in the two last AVEC challenges. However, the pertinence of the CCC as loss function is reassessed in our experiments. CCC is given by equation 1, where x and y are two variables, μ_x and μ_y are their means and σ_x and σ_y are the corresponding

standard deviations. ρ is the correlation coefficient between the two variables.

$$\rho = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

Actually, when the reference is constant over time, ($\sigma_y = 0$), CCC is zero. More generally, when the reference varies slowly, CCC will be almost zero. Consequently, the loss function penalizes conversations where the reference varies slowly ($\sigma_y \simeq 0$), and the trained network will have difficulties to predict correctly such references. We decided to use the root mean square error (RMSE) (see eq. 2 where x_i is a prediction, y_i a reference and n the number of values), as loss function to neutralize this effect.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - x_i\Big)^2} \tag{2}$$

4.2 Cross-corpus experiments: acoustic features and loss functions

Table 3: Average CCC scores on AlloSat and SEWA dev sets with 4 acoustic feature sets and 2 loss functions (*l-ccc* and *l-rmse*). The network is a biLSTM-4. Training and predictions on SEWA has been made on the German conversations.

Features	**AlloSat**		**SEWA**					
	satisfaction		arousal		valence		liking	
	l-ccc	*l-rmse*	*l-ccc*	*l-rmse*	*l-ccc*	*l-rmse*	*l-ccc*	*l-rmse*
eGeMAPS-47	.437	.381	**.487**	.438	**.428**	.404	**.258**	.252
eGeMAPS-88	.564	.514	.316	.201	.237	.211	.119	.077
Mfcc-lib	.675	**.698**	.258	.222	.192	.103	.180	.192
Mfcc-Os	.382	405	.394	.377	.373	.357	.221	.234

In this experiment, the 4 acoustic feature sets described in section 2 are explored as inputs of a biLSTM-4 network. The results are consigned in Table 3. Clearly, and whatever the loss is, eGeMAPS-47 performs better on SEWA dimensions, following by Mfcc-Os. At the other end, Mfcc-lib performs better on AlloSat satisfaction, followed by eGeMAPS-88. This is probably due to the fact that AlloSat contains telephone conversations with diverse background noises, which can alter the extraction of fine-tuned features present in eGeMAPS sets. Moreover, RMSE loss increases performance only when combined with Mfcc features, for satisfaction and liking. Interestingly, these two dimensions are the ones that vary the less according time, consequently have the lowest σ_y. Mfcc-lib with RMSE loss achieve new state-of-art performance on AlloSat.

5 Analysis of the results and post-processing

Even if the score achieved by the best model on satisfaction is high ($ccc = 0.698$), we can observe that predictions vary rapidely with time (see Fig. 2).

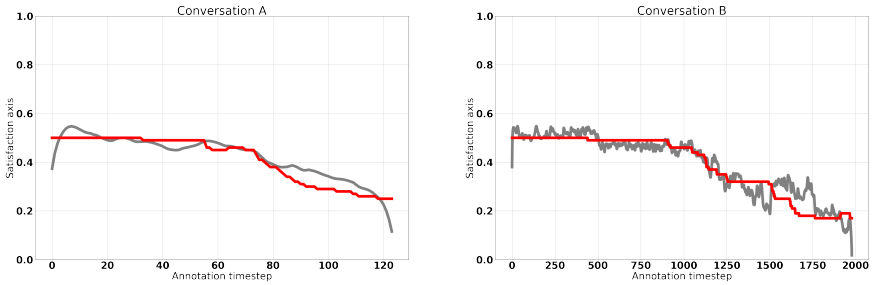


Fig. 2: Evolution of reference satisfaction (red) and its prediction (gray) in two conversations from AlloSat test subset. $ccc(A) = 0.564$, $ccc(B) = 0.903$.

We propose to post treat the prediction with Savistky-Golay smoothing algorithm and polynomial degree of 0. Table 4 confirms that this post-treatment improves the results, achieving new state-of-art results on the satisfaction dimension.

Table 4: CCC scores without and with a smoothing function computed on development and test set of AlloSat on our best model: biLSTM-4, Mfcc-lib features and RMSE loss function.

Features	**Dev**		**Test**	
	Raw	Smoothed	Raw	Smoothed
Mfcc-lib (*l-rmse*)	.698	**.719**	.513	**.570**

6 Conclusion

In this paper, we estimate how much continuous SER is robust to variability issues with cross-corpus experiments.

CNN and RNN were evaluated in the continuous SER task and we conclude that RNNs are robust to cross-corpus conditions and achieve the best results on satisfaction. CNNs perform better on arousal, valence and liking but seem to be very sensitive to filter initialization. We also show that the best feature set (and its implementation) depends on the studied dimension and/or corpus: satisfaction is better represented by MFCCs ($ccc = 0.698$) while arousal, valence and liking are better represented by eGeMAPS-47 features. Indeed, the extraction of fine-tuned features as those in eGeMAPS is probably very sensitive to noisy signals such as telephone. All these results highlight the issue of variability robustness as performances are corpus-dependant. To go further, a post-treatment has been applied on satisfaction predictions showing a significant improvement. The very nice result ($ccc = 0.719$) obtained on AlloSat achieves new state of the art.

References

1. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., et al.: CEICES : Combining efforts for improving automatic classification of emotional user states: a forced co-operation initiative. In: Language and Technologies Conference. pp. 240–245 (2006)
2. Devillers, L., Vaudable, C., Chasatgnol, C.: Real-life emotion-related states detection in call centers: a cross-corpora study. In: Proc. of Interspeech. pp. 2350–2355 (2010)
3. Ekman, P.: Basic Emotions, pp. 301–320. Wiley, New-York (1999)
4. Eyben, F., Batliner, A., Schuller, B.W., Seppi, D., Steidl, S.: Cross-corpus classification of realistic emotions - some pilot experiments. In: LREC 2010 (2010)
5. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., et al.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. on Affect. Computing **7**(2), 190–202 (2016)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (1997)
7. Jing, S., Mao, X., Chen, L.: Prominence features: Effective emotional features for speech emotion recognition. Digital Signal Processing **72**, 216 – 231 (2018)
8. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., et al.: SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. IEEE Trans. on pattern analysis and machine intelligence pp. 1–20 (2019)
9. Lin, L.I.K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**(1), 255–268 (1989)
10. Macary, M., Tahon, M., Estève, Y., Rousseau, A.: AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis. In: Language Resources and Evaluation Conference, LREC 2020 (2020)
11. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schröder, M.: The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. on Affect. Computing **3**(1), 5–17 (2012)
12. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proc. of Int. Conf. on Automatic Face and Gesture Recognition (FG). pp. 1–8 (2013)
13. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., et al.: AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In: Proc. of the 2018 on Audio/Visual Emotion Challenge and Workshop. pp. 3–13. ACM (2018)
14. Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al.: AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In: Proc. of the 9th International on Audio/Visual Emotion Challenge and Workshop. p. 3–12 (2019)
15. Russel, J.: Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective, pp. 295–360. Cambridge University Press, U.K. (1997)
16. Scherer, K.R.: What are emotions ? and how can they be measured ?, chap. Social Science Information, pp. 695–729 (2005)
17. Schmitt, M., Cummins, N., Schuller, B.W.: Continuous Emotion Recognition in Speech - Do We Need Recurrence? In: Proc. Interspeech. pp. 2808–2812 (2019)
18. Tahon, M., Devillers, L.: Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges. IEEE/ACM Transactions on Audio, Speech and Language Processing **24**, 16 – 28 (2016)