Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition

Danqing Luo[1], *Yuexian Zou*[1], *Dongyan Huang*[2]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2] Human Language Technology, Institute for Infocomm Research/A*STAR, Singapore
zouyx@pkusz.edu.cn

Abstract

Speech emotion recognition (SER) is a challenging task due to its difficulty in finding proper representations for emotion embedding in speech. Recently, Convolutional Recurrent Neural Network (CRNN), which is combined by convolution neural network and recurrent neural network, is popular in this field and achieves state-of-art on related corpus. However, most of work on CRNN only utilizes simple spectral information, which is not capable to capture enough emotion characteristics for the SER task. In this work, we investigate two joint representation learning structures based on CRNN aiming at capturing richer emotional information from speech. Cooperating the handcrafted high-level statistic features with CRNN, a two-channel SER system (HSF-CRNN) is developed to jointly learn the emotion-related features with better discriminative property. Furthermore, considering that the time duration of speech segment significantly affects the accuracy of emotion recognition, another two-channel SER system is proposed where CRNN features extracted from different time scale of spectrogram segment are used for joint representation learning. The systems are evaluated over Atypical Affect Challenge of ComParE2018 and IEMOCAP corpus. Experimental results show that our proposed systems outperform the plain CRNN.

Index Terms: speech emotion recognition, convolutional recurrent neural network, joint representation learning

1. Introduction

Perception of emotional states of interlocutors plays an important role in the human social interaction, so does it for human computer interaction. As a fundamental modality of human communication, emotion recognition from speech is particularly important for the computer to analyze the human user's emotional state and respond to it accordingly [1]. In a new era of artificial intelligence, speech emotion recognition (SER) has become an active research area.

Human speech carries complex information. It contains not only linguistic message, but also many attributes of the speaker such as gender, age and identity. How to separate emotion information from them and find proper representations for emotion has been a great challenge for the SER task [2].

SER has been studied since 1980's [3]. The mainstream for speech emotion feature extraction can be classified into two categories. One is the traditional method using handcrafted low-level descriptors (LLDs) and high-level statistic functionals (HSFs) for the feature extraction [4]. LLDs describe speech characteristics in a very short time scale. It mainly contains prosodic, spectral and voice quality features, which are often extracted from short speech segment (such as one to two frames). Specifically, common LLDs include fundamental frequency, energy, formants, zero crossing rate, Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), shimmer, jitter and so on. Compared with LLDs, HSFs describe the dynamic emotional content throughout a whole speech utterance. HSFs are calculated as statistics of LLDs, such as their mean, maximum, minimum, variance, kurtosis, skewness and so on, which essentially represent the global dynamic variation of LLDs. The other method for SER feature extraction is by the use of Deep Neural Network (DNN). With the success of deep neural networks in computer vision and speech recognition, deep learning based SER has drawn a lot of attention [5]. Among various kinds of deep neural networks, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are most commonly used for the SER task and the state-of-art has been achieved on different speech emotion corpus [6] [7]. Research demonstrates that handcrafted features and deep learning features describe the emotion information from different aspects. However, there is still no clear conclusion on the optimal universal feature sets for the SER task.

In this work, we investigate two joint representation learning structures based on CRNN, aiming at capturing richer emotional information from speech. The handcrafted HSFs and CRNN-learned feature describe emotional state of speech from different aspects and complementarity exists between them. Cooperating the handcrafted HSFs with CRNN-learned feature, a two-channel SER system is proposed to jointly learn the emotion-related features with better discriminative property than traditional way. In this system, the two types of features are joined through a hidden layer that projects them into the same representation space. Furthermore, it is noticed that the experimental results of plain CRNN SER system vary a lot with different time durations of the segment, which means the CRNN can learn different contextual emotion representations from various time scales. Keeping this in mind, another two-channel SER system is proposed to capture better contextual characteristics of emotional speech than the previous one. In this system, CRNN features extracted from different time scales of the spectrogram segment are used for joint representation learning.

The performance of systems is evaluated over Atypical Affect Challenge of ComParE2018 and IEMOCAP corpus. Experimental results show that both of our proposed systems outperform the plain CRNN SER system, which confirms their ability to extract robust features for SER task.

10.21437/Interspeech.2018-1832

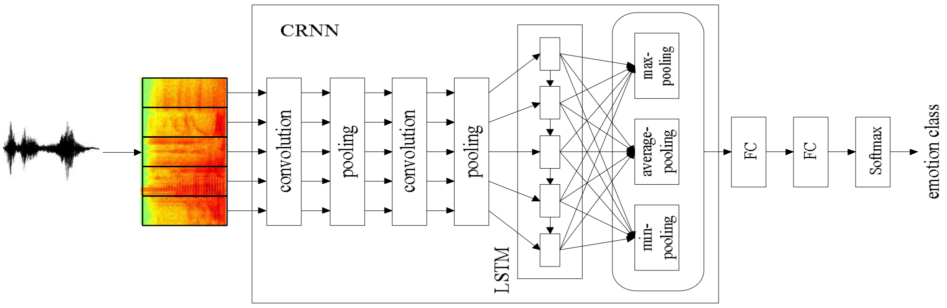

Figure 1: *plain CRNN SER system*

2. Related work

DNN has shown superior ability to learn hierarchical high level representations from raw features and performs well in numerous pattern recognition tasks. In the field of SER, DNN is initially used as a classifier which is substituted for other classifiers, e.g., support vector machine (SVM), which is trained on utterance-level statistical features [8]. In the work of [9], DNN is exploited to extract from short-term LLDs the effective emotional features that are fed into other classifiers for emotion recognition, and makes a breakthrough for SER.

It is clear that the information encoded in speech signals is sequential and the RNN is powerful in dealing with sequential data. As a result, RNN has become a common deep learning structure for SER. For example, in [7], authors proposed a RNN based framework with maximum-likelihood learning criterion to model random label sequence of an utterance, which improves the SER recognition accuracy.

Apart from the RNN, CNN is another popular alternative in the field of SER due to its outstanding performance in image related tasks. Much work has been done for investigating its effectiveness in automatically learning affective representations from spectrogram [10] [11].

Furthermore, there is a raising tendency to combine the CNN with RNN, which is termed as the CRNN model. With the end-to-end training strategy, we can avoid greedily enforcing the distribution of intermediate layers to approximate that of labels, and more proper representations can be obtained from the final layer. A few recent studies on the CRNN have demonstrated its efficiency in SER [12] [13].

3. Proposed method

In this work, the CRNN is used as a base structure to explore the possibility in learning more effective emotional representations from speech. Thus, a plain CRNN SER system is employed as the baseline, upon which two new systems are developed. Firstly, to benefit from the complementarity of handcrafted HSFs, which describe emotional characteristics from another aspect, a two-channel SER system is proposed to learn the joint representation of CRNN-learned feature and HSFs through the neural network. Secondly, to better capture dynamic emotional characteristics of the speech, another two-channel SER system is designed based on the CRNN to integrate features extracted from different time scale of spectrogram segments. The details of abovementioned SER systems are elaborated in the following sections.

3.1. Plain CRNN SER system

Since the CRNN is taken as the main feature extraction network, the SER system composed of plain CRNN is firstly described. On one hand, convolution layers in CNN focus on learning local information of the input. As every point of the feature map derived from convolution layers depends on the multiplication of convolution kernel with the corresponding overlapped region of the input. On the other hand, the RNN with a pooling layer can learn the global representation of the input through complex nonlinear transform between time steps. Therefore, in the SER task, combining CNN and RNN together helps to model subtle local emotion cues and capture contextual emotion information simultaneously.

For the CRNN base structure used in our proposed SER systems, we follow the network proposed by [14] and add more statistic characteristics to the output of RNN by applying the max-pooling and min-pooling. The diagram for plain CRNN SER system is shown in Figure 1.

The network architecture of plain CRNN SER system consists of three parts, where the preceding two parts constitute a CRNN. The first part is a convolutional feature extractor, which takes a spectrogram image representation of a segment of one audio file as input. This feature extractor convolves the input image and pools it in several steps and produces a flattened feature map. Thus, for a speech utterance split into segments in advance, we get CNN-learned feature for every segment. Specifically, for a given utterance denoted as [$\mathbf{s}(1)$, $\mathbf{s}(2)$,…, $\mathbf{s}(T)$], where $\mathbf{s}(t)$ is a segmental spectrogram of original audio input and T is the number of segments, we get a sequence of feature vectors [$\mathbf{c}(1)$, $\mathbf{c}(2)$,…, $\mathbf{c}(T)$] after the CNN module.

The second part is a RNN, where every time step corresponds to a segment of the original audio input. The RNN can handle various length of input, so no audio inputs clipping or padding is applied. The LSTM structure is chosen for the RNN, which is capable of addressing long-term dependencies existing in the sequential data. The LSTM learns from the sequence of CNN feature vectors [$\mathbf{c}(1)$, $\mathbf{c}(2)$,…, $\mathbf{c}(T)$] and also outputs a sequence [$\mathbf{r}(1)$, $\mathbf{r}(2)$,…, $\mathbf{r}(T)$]. Then, similar with HSFs calculated on LLDs, statistics of the LSTM's outputs are computed through pooling layers. In common case, only the average pooling is applied. To get richer statistic information of outputs of LSTM network, we additionally compute maximum pooling and minimum pooling on them and concatenate the resulting pooling vectors $\mathbf{P}_{max}$, $\mathbf{P}_{ave}$, $\mathbf{P}_{min}$ together into one vector. The LSTM is set to have 128 units, thus vector $\mathbf{r}(t)$, $\mathbf{P}_{\text{max}}$, $\mathbf{P}_{\text{ave}}$ and $\mathbf{P}_{\text{min}}$ all have a dimensionality of 128. Let

$\mathbf{r}(t)^i$ represent the i^{th} element of $\mathbf{r}(t)$.The pooling process can be formulated as follows:

$$\mathbf{P}_{max} = (p^1_{max}, \dots, p^{128}_{max}) \text{ , where } p^i_{max} = \max_{1 \le t \le T} \mathbf{r}(t)^i \quad (1)$$

$$\mathbf{P}_{ave} = \sum\nolimits_{1 \le t \le T} r(t) / T \quad (2)$$

$$\mathbf{P}_{min} = (p^1_{min}, \dots, p^{128}_{min}) \text{ , where } p^i_{min} = \min_{1 \le t \le T} \mathbf{r}(t)^i \quad (3)$$

Besides, the subsequent third part of the network comprises two fully-connected layers and one softmax layer that predicts the emotion category. Their size is 128, 32 and 4, respectively.

3.2. Our proposed HSF-CRNN SER system

As we stated before, handcrafted HSFs characterize statistic features of various prosodic, spectral and voice quality features while CRNN only learns from the spectrogram. The two types of features describe emotional state of speech from different aspects and lie in respective feature space. Therefore, the complementarity exists between these two kinds of features. To fully utilize emotional information encoded in speech and extract robust features for SER, the two types of features can be cooperated, which expects better performance. Research has demonstrated that DNN can effectively extract discriminative features that approximate the non-linear dependencies between features in the original set [15]. Hence, the neural network is used to construct a joint representation of handcrafted HSFs and CRNN learned feature. Specifically, through a hidden layer, the two types of features are projected together into the same feature space while dimensionality of the original feature is reduced.

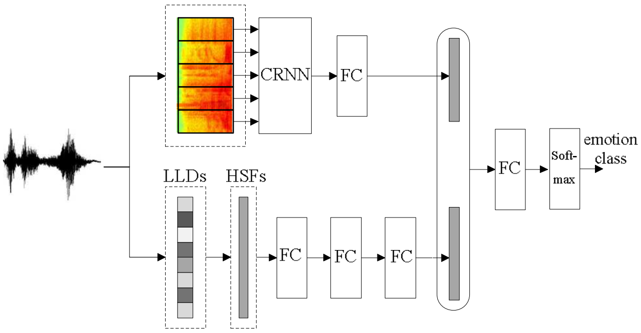

Figure 2: *HSF-CRNN SER system*

In our work, a two-channel SER system is proposed to realize the joint representation learning structure. The diagram is shown in Figure 2.

In this SER system, given a speech utterance, it is firstly processed in two parallel channels. In one channel, the spectrogram is computed and then divided into segments. These segments are input to a CRNN network. The CRNN outputs a concatenated pooling vector of size 384 and then one hidden layer of size 128 follows after it. In another channel, the raw waveform is segmented into frames. LLDs are extracted from utterance frames and HSFs are further computed. Three hidden layers map the high-dimensional HSFs successively into feature space of lower dimensions through non-linear transform and output a feature vector in the end. The two feature vectors output from respective channel are concatenated and used as the input to next hidden layer that projects them into a joint feature space. Next, the joint feature representation is passed to one softmax layer to make the classification. The whole system is trained in an end-to-end fashion, learning both to extract joint feature representation and to perform SER task.

3.3. Our proposed Multi-CRNN SER system

In our preliminary experiments on plain CRNN SER system, we found that the time duration of divided spectrogram segment significantly influences the accuracy of emotion recognition, which means that the CRNN can learn different contextual emotion information from different time scale of the spectrogram. Intuitively, the longer duration one spectrogram segment has, the more global feature CRNN can learn from the entire utterance. Meanwhile, due to the multilayer nature of DNN, the deeper layer is hypothesized to represent the raw data in a more abstract way. Motivated by this, we study another joint representation learning structure based on CRNN.

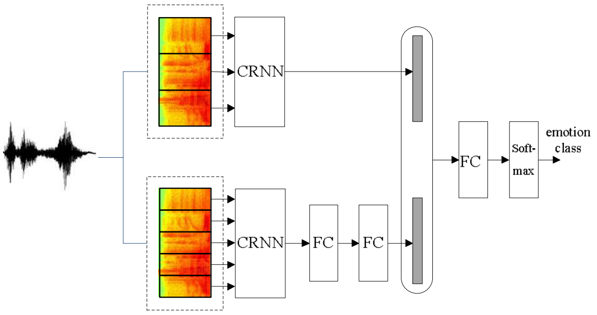

Figure 3: *multi-CRNN SER system*

In this structure, we investigate to add CRNN-learned feature extracted from the longer time scale of spectrogram segment to the deeper layer of a neural network. And joint representation of two CRNN-learned features corresponding to different time scales is extracted through a hidden layer. Diagram of this SER system is shown in Figure 3.

Given a speech utterance, the spectrogram is firstly computed. Then, two sliding windows of different length are applied to the spectrogram respectively, to generate segments of two time scales. Deep features are extracted for each type of spectrogram segments with a CRNN. For the CRNN feature learned from the shorter segment and hence the smaller time scale, two hidden layers of size 128 are employed to learn more abstract representation. Output vector of the last hidden layer is concatenated with CRNN feature learned from the longer segment. Next, a hidden layer of size 128 projects the features concatenated by CRNN features corresponding to different time scales together into one feature space. Finally, one softmax layer is applied to complete the SER task. Similarly, this SER system is also trained in an end-to-end manner.

4. Experiments

We evaluate the above three SER systems over Atypical Affect Challenge of ComParE2018 and IEMOCAP corpus. Experiments show that the two SER systems with joint representation learning structure both outperform plain CRNN SER system.

4.1. Data

The goal of Atypical Affect Challenge is to classify emotion category from angry, happy, sad and neutral. Provided data is partial audio tracks of the EmotAsS database [16], which contains speech of cognitively impaired subjects. Data has been partitioned into training set, develop set and test set in advance. The whole dataset is extremely unbalanced with much more utterances of neutral category than other three categories.

IEMOCAP database [17] is composed of five sessions from 10 speakers. In each session a pair of actors talk to each other according to given scripts or improvise in a pre-defined situation. We take the same four emotion categories as Atypical Affect Challenge from this database. There are totally 5498 utterances in the extracted dataset, with 1090, 1627, 1704 and 1077 utterances belonging to angry, happy, neutral and sad category, respectively. The experiments are conducted in a speaker-independent manner. Four sessions from 8 speakers are chosen as training data, the remaining one from 2 other speakers as test.

4.2. Experimental setup

In all the experiments, input audio signals are converted into frames using a 25-ms window sliding at 10-ms each time. Log-mel spectrogram is used as the input to CRNN. Firstly, 1024-point short time Fourier transformation is applied to Hamming windowed frame, then 40 log-mel filter banks are applied to compute log-mel spectrogram. For plain CRNN, the size of one spectrogram segment is set to 30 frames and every two adjacent segments has an overlap of 10 frames. Thus, the length of a segment is 315 ms. Research has shown that a speech segment longer than 250 ms can encode sufficient emotion information [18]. For another channel in multi-CRNN SER system, the size of one spectrogram segment is set to 50 frames and the overlap is 25 frames. The corresponding length of a segment is 515 ms.

The CRNN base structure contains two consecutive convolutions with maxpooling following each. Research shows that the full-spectrogram temporal convolution is more favorable for SER [19]. Filter of the full-spectrogram temporal convolution covers the entire spectrogram axis and convolves with input tensor only in the temporal direction. Hence the first convolution layer of CRNN base structure has 30 filters of size 40×5. The second convolution layer is set to have 30 filters of size 1×3. Both convolution layers are activated by Relu. The two maxpooling layers have a same pooling size of 1×3 and a stride of 2.

For the HSF-CRNN SER system, the LLDs and HSFs provided for ComParE2016 challenge are used. The final HSFs contain 6373 features. In the HSF channel, the three hidden fully connected (FC) layers have size of 1024, 520 and 128, respectively. The joint learning FC layer has 32 nodes and the softmax layer has 4 nodes. In the three SER systems, all FC layers are activated by sigmoid function and all networks are trained using the cross entropy as the objective function.

4.3. Experimental results

To measure the performance, the weighted accuracy (WA) and unweighted accuracy (UA) are reported on IEMOCAP corpus. Due to the unbalanced distribution of EmotAsS dataset, UA is used as the primary criterion for Atypical Affect Challenge.

A summary of the SER results given by different systems on IEMOCAP corpus is presented in Table 1. The multi-CRNN SER system achieves WA of 57.53% and UA of 58.06%, which are both higher than the plain CRNN system. Furthermore, the HSF-CRNN system improves the accuracy to 60.35% and 63.98% for WA and UA, outperforming the plain CRNN system by around 3.8% and 7.6% in terms of WA and UA. Results on the develop set of Atypical Affect Challenge can be seen in Table 2. Compared with plain CRNN SER system, similar improvement from the multi-CRNN and HSF-CRNN system is observed, where the highest UA is achieved by HSF-CRNN system with 32.45%. Therefore, it can be concluded that the SER system with joint learning structure learns more robust emotional features than the plain CRNN system. Moreover, experimental result suggests that between the two different joint learning structures, combining HSFs with CRNN-learned feature is superior to joining CRNN-learned features extracted from different time scale of spectrogram.

We submit the emotion recognition result on the test set of Atypical Affect Challenge predicted by HSF-CRNN system, and get the UA of 35.795%. The test result is higher than the best one of provided baseline using end-to-end CNN-LSTM network, which is 28.0% [20]. It indicates that although the baseline is fine-tuned on develop set with UA of 41.8% as shown in Table 2, it doesn't learn robust enough emotional feature for the task of SER, which results in a low accuracy on test set. Compared with the baseline, our proposed system performs better on unseen data, which proves its ability in robust feature extraction for SER.

Table 1: *SER results of three systems on IEMOCAP.*

SER system	WA (%)	UA (%)
Plain CRNN	56.54	56.42
Multi-CRNN	57.53	58.06
HSF-CRNN	60.35	63.98

Table 2: *UA (%) of three systems on Atypical Affect Challenge.*

SER system	Develop set	Test set
Plain CRNN	29.37	-
Multi-CRNN	30.65	-
HSF-CRNN	32.45	35.795
*Baseline[20]	41.8	28.0

5. Conclusions

In this work, we study on SER task under CRNN framework and two novel SER systems have been developed. Our main considerations lie on the joint representation learning strategy to fully make use of the hand-crafted features as well as the multi-scale information from speech spectrums. Our experimental results show that our proposed SER systems outperform the baseline plain CRNN SER system. The results indicate the efficiency of joint learning structure to extract robust emotion representation from handcrafted features and CRNN-learned spectrogram features, where richer emotional information can be captured for SER task. In our future work, we will study the contribution of different types of feature to SER task and explore other representation learning method for SER.

6. Acknowledgements

This work was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ 20170306165153653 & JCYJ20170817160058246).

7. References

[1] Ramakrishnan, S., and Ibrahiem MM El Emary. "Speech emotion recognition approaches in human computer interaction." Telecommunication Systems 52.3 (2013): 1467-1478.

[2] Schuller, Björn, et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge." Speech Communication 53.9-10 (2011): 1062-1087.

[3] Van Bezooijen, Renée, Stanley A. Otto, and Thomas A. Heenan. "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics." Journal of Cross-Cultural Psychology 14.4 (1983): 387-406.

[4] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, and Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." Artificial Intelligence Review 43.2 (2015): 155-177.

[5] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

[6] Niu, Yafeng, et al. "A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks." arXiv preprint arXiv:1707.09917 (2017).

[7] Lee, Jinkyu, and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." (2015).

[8] Stuhlsatz, André, et al. "Deep neural networks for acoustic emotion recognition: raising the benchmarks." Acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on. IEEE, 2011.

[9] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[10] Anand, Namrata, and Prateek Verma. "Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data." Technical Report. Stanford University, 2015.

[11] Mao, Qirong, et al. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE Transactions on Multimedia 16.8 (2014): 2203-2213.

[12] Huang, Che-Wei, and Shrikanth Shri Narayanan. "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition." Multimedia and Expo (ICME), 2017 IEEE International Conference on. IEEE, 2017.

[13] Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific. IEEE, 2016.

[14] Huang, Che-Wei, and Shrikanth Narayanan. "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition." arXiv preprint arXiv:1706.02901 (2017).

[15] Kim, Yelin, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[16] Hantke, Simone, et al. "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings." Proc. Interspeech 2017 (2017): 3137-3141.

[17] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42.4 (2008): 335.

[18] Kim, Yelin, and Emily Mower Provost. "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[19] Anand, Namrata, and Prateek Verma. "Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data." Technical Report. Stanford University, 2015.

[20] Björn W. Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian Pokorny, Eva-Maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, Stefanos Zafeiriou: "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats", Proceedings INTERSPEECH 2018, ISCA, Hyderabad, India, 2018.