CONTEXT-AWARE ATTENTION MECHANISM FOR SPEECH EMOTION RECOGNITION

Gaetan Ramet[*‡], *Philip N. Garner*[†], *Michael Baeriswyl*[‡], *Alexandros Lazaridis*[‡]

[*] Ecole Polytechnique Federale de Lausanne, Switzerland
[†] Idiap Research Institute, Martigny, Switzerland
[‡] Artificial Intelligence and Machine Learning Group, Swisscom
gaetan.ramet@epfl.ch, phil.garner@idiap.ch, {michael.baeriswyl, alexandros.lazaridis}@swisscom.com

ABSTRACT

In this work, we study the use of attention mechanisms to enhance the performance of the state-of-the-art deep learning model in Speech Emotion Recognition (SER). We introduce a new Long Short-Term Memory (LSTM)-based neural network attention model which is able to take into account the temporal information in speech during the computation of the attention vector. The proposed LSTM-based model is evaluated on the IEMOCAP dataset using a 5-fold cross-validation scheme and achieved 68.8% weighted accuracy on 4 classes, which outperforms the state-of-the-art models.

Index Terms— Speech Emotion Recognition, Attention, Deep Learning, Neural Network

1. INTRODUCTION

As automatic speech recognition and synthesis become more and more ubiquitous, the possibility of recognizing and synthesizing even higher level semantics opens up. Emotion (or affect) is one such semantic. For example, in the context of a speech-to-speech translation system, we would like to be able to reflect the fact that a speaker is happy or angry to the translated voice. In the context of a dialogue agent in a call center, we would like to know if the client is satisfied with the process of their query; a client becoming angry or frustrated may be a cue to transfer them to a human operator. To this end, and in the context of this paper, we are interested in general in the recognition of emotion in speech.

Speech emotion recognition (SER) is the process of automatically determining the underlying emotional state of a person from a sample of its speech. For many years, the state-of-the-art SER models were developed using statistical parametric features followed by a classification or regression algorithm [1]. Nonetheless, over the last years due to the increase of available data and computational power, deep learning models such as Neural Networks (NNs) have managed to outperform the traditional machine learning algorithms. One particular architecture which has emerged is based on the combination of Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs). This approach is able to exploit the temporal information in speech to predict utterance-level emotion labels.

Neural network based attention mechanisms are widely used over the recent few years in deep learning. These techniques were firstly introduced in the Neural Machine Translation field by Badhanau et. al. [2]. Ever since, attention mechanisms have been applied to various machine learning tasks, such as key-term extraction [3], image classification [4], semantic segmentation [5] and also recently to SER [6].

In its simplest form, an attention mechanism can be described as a single vector. Let's define $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_T]$ as an input matrix of shape $T \times F$, with T the number of frames and F the number of features per frame. In this case, the attention model is denoted as a weight vector $\boldsymbol{w} \in \mathbb{R}^F$, which is multiplied with $\boldsymbol{H}$:

$$\boldsymbol{M} = \boldsymbol{w}^T \boldsymbol{H} \tag{1}$$

M is denoted as the "attention map". This map can be used in various ways. The most common use is normalizing it to unity and using it as a weighting factor for the RNN output, to perform a weighted pooling. One possible way to do that is by using a softmax:

$$\boldsymbol{R} = \boldsymbol{H}\boldsymbol{\alpha}^T, \tag{2}$$

where,

$$\alpha_i = \text{softmax}(\boldsymbol{m}_i) \tag{3}$$

$$= \frac{\exp(\boldsymbol{w}^T \boldsymbol{h}_i)}{\sum_t \exp(\boldsymbol{w}^T \boldsymbol{h}_t)} \tag{4}$$

The main idea behind attention mechanisms is to "help" the model to learn where to "look for" the information that is meaningful for the task at work. It can then focus on the relevant parts while disregarding noisy or irrelevant data. The attention mechanism can take multiple forms. A simple model

P. Garner received funding from the EU H2020 project Scalable Understanding of Multilingual Media (SUMMA) project no. 688139. http://summa-project.eu/

such as the one presented in Eq.1 and 2 was successfully used by Zhou et al. [7] for relation classification. A more complex model was developed by Yang et al. [8] where an additional fully-connected layer is used to compute $\boldsymbol{H}$. The common feature across different attention mechanisms is the fact that they are almost always trained along with the rest of the network by back-propagation.

In order to improve the performance of the state-of-the-art models for SER, we investigate the use of different attention mechanisms. The goal is to localize the parts of speech conveying emotional load in the utterance and weight the different frames based on their emotional relevance. We hypothesize that the attention mechanism will be able to localize the meaningful information, while disregarding the noisy data. However, the majority the attention models introduced in the literature are applied on a frame-by-frame level. This makes them incapable of taking into account the contextual information of the neighboring frames. To alleviate this problem, we designed a new attention model based on a bi-directional LSTM (BiLSTM), which would be able to exploit the sequentiality in speech. We hypothesize that the ability of LSTMs to model the context of neighboring frames will create more robust attention vectors.

The rest of this paper is organized as follows: the proposed method with the LSTM attention model we investigate is introduced in Section 2. The experimental setup along with the other attention mechanisms used for comparison are described in Section 3. The results are presented and analyzed in Section 4. In Section 5 the final conclusions are presented.

2. PROPOSED METHOD

We introduce a new attention mechanism which relies on the bi-directional long short-term memory (BiLSTM)'s ability to model the temporal dependencies of sequential data such as speech. An LSTM unit takes as input a feature vector $\boldsymbol{x}_t$ and computes the output vector $\boldsymbol{h}_t$ using the previous output vector $\boldsymbol{h}_{t-1}$ and the previous memory cell state $\boldsymbol{c}_{t-1}$ as follows:

$$\boldsymbol{f}_t = \sigma_g(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \quad (5)$$

$$\boldsymbol{i}_t = \sigma_g(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \quad (6)$$

$$\boldsymbol{o}_t = \sigma_g(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad (7)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} \boldsymbol{i}_t \circ \sigma_c(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \quad (8)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \circ \sigma_h(\boldsymbol{c}_t) \quad (9)$$

with σ_g the sigmoid function, σ_c and σ_h the hyperbolic tangent when no peephole connection is used. Thanks to its specific design, the LSTM cell is able to retain information over time, incorporating new elements and discarding irrelevant ones, as the input flows. This property makes the LSTM cell especially suitable for tasks such as SER where the sequentiality of the data is important.

We chose to use a bi-directional LSTM in our novel attention mechanism as there is a strong time dependency on the parts of speech that conveys emotions and we want to take in account both the past and future contexts. The proposed model uses the BiLSTM to compute the matrix $\boldsymbol{H}$ which is then processed as described in Eq. 1-2. We decided to apply a sigmoid activation instead of the usual softmax normalization to compute the attention vector $\boldsymbol{\alpha}$, as the softmax forces a very low number of frames to have high activation. By using a sigmoid activation instead, we can ensure that many frames get a high activation level aiming for an overall smoother attention vector.

Using a BiLSTM inside the attention module allows for the model to take into account each frame in its context within the full utterance. In this way, we hypothesize that the frames are context dependent; determining the level of attention a frame should receive does not depend only on the frame itself, but also on its past and future context. The use of a BiLSTM also allows us to control the complexity of the model by varying the output size of the LSTM.

With this kind of attention mechanism, we can force the model to learn meaningful features as well as to localize the emotionally salient parts of an utterance. The attention vector that is generated will not only give low weights to the silent parts of an utterance, in a VAD fashion, but also to the non-emotional parts which might be voiced. Moreover, by incorporating an LSTM inside the attention mechanism, we expect the model to cope better with the sequentiality of the data to compute the attention vector.

3. EXPERIMENTAL SETUP

We investigate the benefits of using attention mechanisms for SER by applying different attention mechanisms on top of a state-of-the-art baseline model. The baseline model is inspired from the one in [6], which is a DNN-RNN. We extract a frame-level feature vector using the openSMILE toolkit [9], which we feed as input to the neural network. Features are extracted using a sliding window of 25ms length, with a shift of 10ms. We extract a 32-dimensional feature vector per frame, which is composed of hand-crafted Low-Level descriptors (LLDs) often used for SER, namely pitch, energy, zero-crossing rate, voicing probability, 12 mel-frequency cepstral coefficients (MFCCs) as well as the first derivative of each of these quantities.

The model is composed of 3 ReLU-activated fully-connected layers with 256 nodes each, followed by a bi-directional LSTM with 128 units and tanh activation. The outputs of the RNN are concatenated and a mean temporal pooling operation is applied. Finally, the resulting 256-features vector is fed to a softmax classifier, which in our case is a fully-connected layer with 4 outputs and softmax activation since we work on a 4-class problem.

To this baseline model, we applied different kind of at-

Table 1: Performance of different attention mechanisms in terms of Weighted Average (WA) and Unweighted Average (UA) for full dataset and improvised dataset

Attention Model	Full dataset		Improvised speech only	
	WA	**UA**	**WA**	**UA**
Baseline (mean pool)	60.5	58.3	62.2	63.3
Zhou et al. [7]	60.7	58.7	64.7	62.4
Yang et al. [8]	61.2	58.4	64.8	64.1
Proposed model	**62.5**	**59.6**	**68.8**	**63.7**

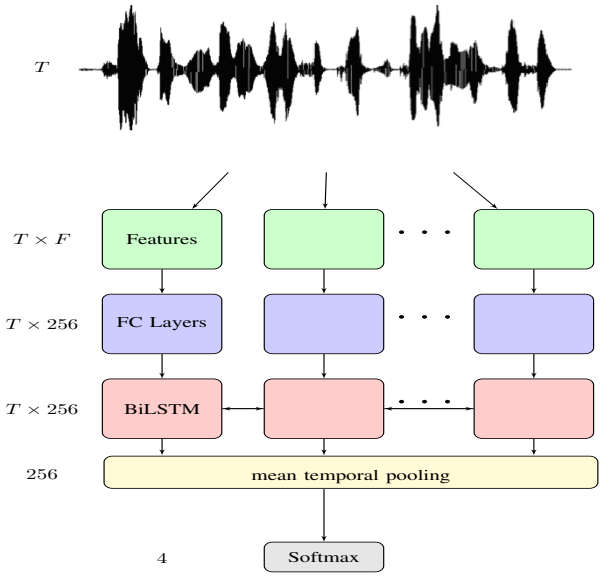


Fig. 1: Architecture of the baseline model. The output size of each part is shown on the left

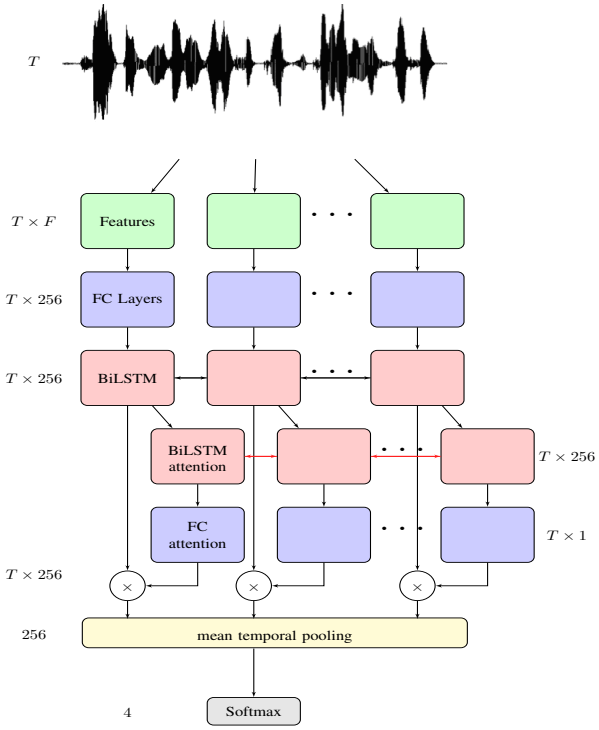


Fig. 2: Architecture of our proposed model, with the LSTM attention mechanism

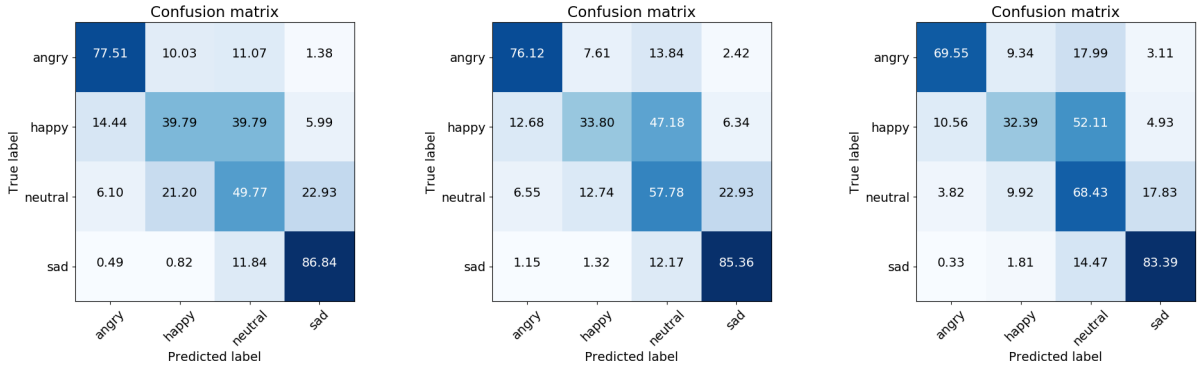
tention mechanisms, with different levels of complexity and different behavior. In this work, we evaluate the performance of 4 different strategies: first a simple mean pooling strategy with no attention mechanism, then two existing attention mechanisms introduced in the literature by Zhou et. al [7] and Yang et. al. [8] respectively, and finally, we also evaluate our proposed attention model.

The three attention models are applied on top of the BiLSTM. The attention model in Zhou et al. [7] consists of a single attention vector $\boldsymbol{w}$ used as described in Eq. 1 and 2 and takes as input the output of the BiLSTM directly. The attention model of Yang et al. [8] passes the output of the BiLSTM through a fully-connected layer to generate the matrix $\boldsymbol{H}$. The fully-connected layer can have any output size, which gives some ability to control the complexity of the attention model. The proposed attention model uses a second BiLSTM layer to generate the feature matrix $\boldsymbol{H}$, as described in 2. The architecture of the proposed attention model is depicted in Figure 2.

All experiments were done on the IEMOCAP database (Interactive Emotional Dyadic Motion Capture) [15]. IEMOCAP is one of the largest available dataset for SER. It contains approximately 12 hours of spoken content, split in 5 sessions with 2 different speakers each, for a total of 10 speakers. In each session, the two professional actors performed scripted and improvised scenarios. As it was shown in [14], the improvised part of the dataset could be characterized as an easier task for prediction. In order to present a more detailed comparison with state-of-the-art work on this database, we decided to tackle the problem using both the full dataset and the improvised speech.

Table 2: Performance of SER models in terms of Weighted Average (WA) and Unweighted Average (UA) for full dataset and improvised dataset [1]

Models (CV scheme)	Full dataset		Improvised speech only	
	WA	**UA**	**WA**	**UA**
Mirsamamdi et al. (5-fold CV) [6]	63.5	58.8	—-	—-
Etienne et al. (10-fold CV) [10]	—-	—-	64.5	61.7
Tzinis et al. (5-fold CV) [11]	—-	—-	64.2	60.0
Huang et al. (Leave-one-session-out) [12]	59.4	50.0	—-	—-
Lee et al. (5-fold CV) [13]	—-	—-	62.9	63.9
Neumann et al. (5-fold CV) [14]	56.1	—-	62.1	—-
Proposed model (5-fold CV)	**62.5**	**59.6**	**68.8**	**63.7**



(a) Baseline (b) Yang et al. [8] (c) Proposed model

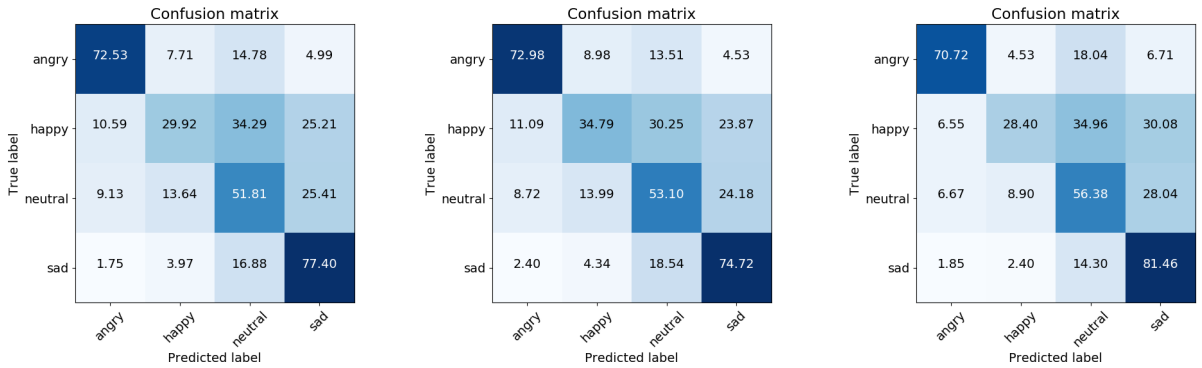
Fig. 3: Confusion matrix of the baseline, Yang et al. model [8] and the proposed model on the improvised set

As there is no predefined train/test split for IEMOCAP, we decided to apply a 5-fold cross-validation scheme using 4 sessions as training set and the last session as testing set to ensure speaker independence. We report the average accuracy over the 5 cross-validation folds, which is our final performance measurement. To stay consistent with most of the literature on IEMOCAP, we keep only the samples belonging to the 4 classes *angry*, *happy*, *neutral* and *sad*, without merging the *happy* and *excited* classes as done in [14]. Moreover, we present both Weighted Accuracy (WA) and Unweighted Accuracy (UA) as the classes are not uniformly distributed. We propose an evaluation procedure as complete, transparent and easily reproducible as possible whilst retaining comparability with other published results on the same dataset, i.e. IEMOCAP, to allow for a fair comparison with the literature to come.

Due to the use of an RNN architecture in the proposed network, a fixed number of frames is required for the training samples. Following the insights of [14], the fixed number of frames was set to 500, for a 5 seconds long utterance. We crop longer utterances and pad shorter ones with zeros to have equal length utterances.

Before feeding the feature vectors to the network, features are normalized by the mean and standard deviation of the neutral speech of the training set, which differs for each split. Moreover, to further avoid overfitting and improve the model's adaptability to unseen data, we augment the data by adding random white noise with a variance $\sigma^2 = 0.4$.

We built the proposed model using Tensorflow [16] and train it by back-propagation using an Adam optimizer with learning rate of 3×10^{-5} while feeding mini-batches of 32 utterances. The entire network is regularized with l_2-regularization with a factor $\gamma = 5e-2$ and dropout is applied on all layers apart from the attention model with a keep probability of 0.9. The values for these hyperparameters were chosen as the ones yielding the best performance on the average of the 5 cross-validation splits. To counter the class-imbalance in each split, we weight each sample of the class c in the loss function by a factor $w_c = \frac{N_{tot}}{N_{classes} N_c}$. We trained each cross-validation split for 200 epochs and keep the best performing model in terms of summed UA and WA, to ensure good performances on the whole dataset and on each separate class.



(a) Baseline (b) Yang et al. [8] (c) Proposed model

Fig. 4: Confusion matrix of the baseline, Yang et al. model [8] and the proposed model on the full dataset

4. EXPERIMENTAL RESULTS

The results of our experiments are presented in Table 1. The LSTM-based attention model outperforms the baseline model and the standard attention models in both WA and UA on the full dataset and in WA on the improvised set. By using our LSTM-based model, a 2% absolute improvement in WA over the full dataset and a 6.6% absolute improvement on the improvised set, were achieved compared to the baseline model. This can be explained by the nature of the LSTM-based attention model ; due to its ability to cope with the contextual information in speech data, using an LSTM to compute the attention vector is shown to be a better suited solution for SER. This increase in performance validates our hypothesis that contextual information is important in the computation of the attention vector.

The LSTM-based attention model performed better on the *neutral* and *sad* class than two other attention models, as can be seen on the confusion matrix presented in Figure 3 and 4. This can be explained by the nature of these two emotions; *neutral* and *sad*, as opposed to *happy* and *angry* are "slow"-paced emotions, less exclamatory and explicit and more passive [17]. Due to this fact, the emotional load is spread more widely in *sad* and *neutral* utterances, which makes an LSTM-based attention model better suited to detect them. On the other hand, angry and happy emotions are more "fast"-paced, energetic and active, and so, much more localized in time.

Comparison of the performance of the proposed model with the literature can be found in Table 2. We were able to compare the proposed model to all the previous work on IEMOCAP, whether the improvised set only or the full dataset were used. The proposed model outperforms the state-of-the-art models on both the improvised partition and the full IEMOCAP dataset, in terms of WA and UA.

[1]Results from [11, 12, 13, 14] were rounded to one decimal digit.

5. CONCLUSIONS

We evaluated the performance of various attention-based models applied to the state-of-the-art speech emotion recognition deep learning based systems. We showed that using attention mechanisms can yield improvement in the accuracy by focusing on emotionally relevant parts of the speech input. Moreover, we introduced a new LSTM-based attention model which can learn in a more robust way to localize the emotionally salient part of speech by taking into account the sequentiality of the speech data. The improvements in performance given by this new LSTM-attention model shows that the context of a given frame is important to compute the attention vector. Additionally, a standardized evaluation procedure for benchmarking models on the IEMOCAP dataset was introduced and showed that the proposed model constitutes as the new state-of-the-art model, with respect to both the full dataset and the improvised part only.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] Sheng-syun Shen and Hung-yi Lee, "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection," *arXiv preprint arXiv:1604.00077*, 2016.

[4] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xi-

aoou Tang, "Residual attention network for image classification," *arXiv preprint arXiv:1704.06904*, 2017.

[5] Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, and Alex C Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *arXiv preprint arXiv:1803.02563*, 2018.

[6] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[7] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, vol. 2, pp. 207–212.

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[10] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch, "Speech emotion recognition with data augmentation and layer-wise learning rate adjustment," *arXiv preprint arXiv:1802.05630*, 2018.

[11] Efthymios Tzinis and Alexandras Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 190–195.

[12] Che-Wei Huang and Shrikanth S Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition.," in *INTERSPEECH*, 2016, pp. 1387–1391.

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

[14] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[17] Nico H Frijda, *The emotions*, Cambridge University Press, 1986.