INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Learning Utterance-level Representations with Label Smoothing for Speech Emotion Recognition

Jian Huang[1,3]*,Jianhua Tao*[1,2,3]*,Bin Liu*[1]*,Zheng Lian*[1,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{jian.huang, jhtao, liubin, zheng.lian}@nlpr.ia.ac.cn

Abstract

Emotion is high-level paralinguistic information characteristics in speech. The most essential part of speech emotion recognition is to generate robust utterance-level emotional feature representations. The commonly used approaches are pooling methods based on various models, which may lead to the loss of detailed information for emotion classification. In this paper, we utilize the NetVLAD as trainable discriminative clustering to aggregate frame-level descriptors into a single utterance-level vector. In addition, to relieve the influence of imbalanced emotional classes, we utilize unigram label smoothing with prior emotional class distribution to regularize the model. Our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database reveal that our proposed methods are beneficial to performance improvement, which is 3% better than other models.

Index Terms: speech emotion recognition, NetVLAD, unigram label smoothing

1. Introduction

Emotions conveys underlying intent of speech signals, which can help intelligent human-machine interaction systems to understand the users' intentions [1]. Emotions can be quantified with discrete categories statically over utterances [2]. One challenge arising is to obtain robust utterance-level feature representations for emotion classification.

The researchers have proposed many methods to handle this problem. Traditionally, the emotional speech is encoded into one feature vector which is the statistics of low-level frame-based handcrafted acoustic features [3]. However, these is no still consensus about appropriate emotional acoustic features. Another type of approaches are pooling methods based on various models which convert emotional temporal sequence to a fixed dimension feature vector. Han et al. [4] constructed utterance-level features from segment-level probability distributions using deep neural networks. Tzinis et al. [5] introduced recurrent neural networks to capture emotional temporal information, which only considered final state of the recurrent layers. Chao et al. [6] compared different pooling methods and highlighted the advantage of mean pooling. Mirsamadi et al. [7] focused on specific regions of speech signal with attention weights that were more emotionally salient, which provided more accurate predictions. Huang et al. [8] extracted discriminative emotional embedding features based on a triplet framework, which also contained pooling operations. However, these pooling methods would loss dynamic temporal information that strongly reflects a change in emotional state [9].

Different from these methods, we utilize a trainable generalized Vector of Locally Aggregated Descriptors (NetVLAD) to settle this problem. VLAD [10] stored the sum of residuals vector between the descriptors and cluster centers to produce the feature vector. Arandjelovic et al. [11] replaced the hard assignment of VLAD with soft assignment, namely NetVLAD, to make the VLAD pooling differentiable for neural networks. NetVLAD would generate more discriminative representations based on different cluster centers for speech emotion recognition, which has been shown to outperform pooling methods in the place recognition [11] and speaker recognition [12].

The effectiveness of speech emotion recognition depends on the quality of utterance-level feature representations, and is also affected by the distribution of emotional classes. However, speech emotion recognition usually encounters the problem of imbalanced emotional classes. For instance, the number of neutral and happy is usually more than other classes. Actually, this situation also exists in real life. Wang et al. [13] collected speech emotion utterances from a Microsoft spoken dialogue system and found similar imbalanced phenomenon. Many users were excited to talk with the dialogue system, so there were lots of happy samples. There were only a small number of sad utterances in the dataset because people did not want to talk with a chatbot when they were in a sad mood. In a word, imbalanced emotional classes is reasonable from a practical perspective.

However, the researches have shown that the class imbalance will cause performance degradation since the class owning majority samples would affect the learning behavior of the deep neural networks by dominating their gradient [14]. As a result, the model would favor the class owning more training samples. Some simple methods such as down-sampling and data-generating can relieve the influence [15], which are suboptimal. Zhang et al. [16] introduced the focal loss to handle the problem of imbalanced emotional classes, which suppressed the contribution of majority samples and gained more focus on the minority samples. Li et al. [17] replaced the angular softmax with softmax to alleviate the severe data imbalance. Another strategies are to penalize the entropy of network's output distribution with label smoothing regularization [18]. Label smoothing reduces overfitting by preventing a network from assigning full probability to each training example. Further, Pereyra et al. [19] smoothed the labels with data's own distribution, defined as unigram label smoothing, to improve state-of-the-art models. In this paper, we use label smoothing to restrain the problem of imbalanced emotional classes.

In this paper, we utilize the NetVLAD to generate effective utterance-level feature representations for speech emotion recognition. Besides, label smoothing is used to relieve the

http://dx.doi.org/10.21437/Interspeech.2020-1391

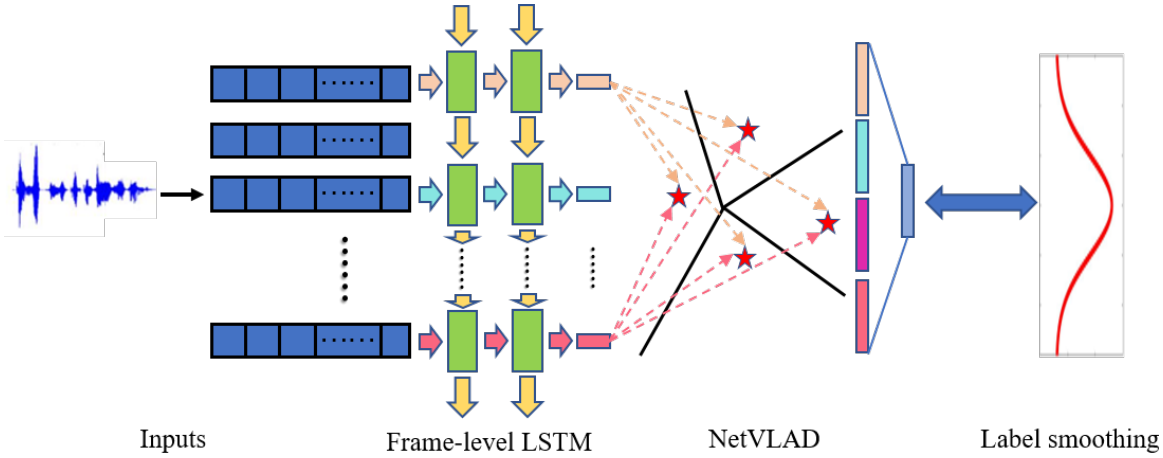


Figure 1: *The overview of our proposed model including three parts: frame-level emotional LSTM model, the NetVLAD layer and unigram label smoothing.*

problem of imbalanced emotional classes. The rest of the paper is organized as follows: section 2 introduces the proposed methods. Section 3 presents the database and acoustic emotional features. In section 4, we describe experimental results and analysis. Finally, we conclude the paper in section 5.

2. Proposed methods

Our proposed speech emotion recognition model is shown in the Figure 1 including frame-level emotional LSTM layer, the NetVLAD layer and label smoothing. The inputs of the model are low-level frame-based features which will be described in the section 3.2. The middle frame-level LSTM layer is responsible to encode emotional dynamic temporal information. The next NetVLAD layer aggregates the outputs of LSTM layer along the temporal axis to produce fixed-length high-level representations followed by a softmax classifier. Finally, we utilize label smoothing to penalize low entropy output distributions.

2.1. NetVLAD

We utilize the NetVLAD to produce high-level feature representations. Intuitively, the VLAD layer can be thought as trainable discriminative clustering: every frame-level descriptor will be softly assigned to different clusters, and their residuals are encoded as the NetVLAD vector outputs. The NetVLAD layer takes dense descriptors from LSTM sequences as inputs in Figure 2 and produces a single $K \times D$ matrix V, where K refers to the number of chosen clusters, and D refers to the dimension of each cluster. Concretely, the matrix of descriptors V is computed using the following equation:

$$V(k,j) = \sum_{t=1}^{T} \frac{e^{w_k x_t + b_k}}{\sum_{k'=1}^{K} e^{w_{k'} x_t + b_{k'}}} \left(x_t(j) - c_k(j)\right) \quad (1)$$

where $\{w_k\}$, $\{b_k\}$ and $\{c_k\}$ are trainable parameters with $k \in [1, 2, ..., K]$, T is the frame length of speech samples.

The first term of (1) corresponds to the soft assignment weight of the input vector x_i for cluster k, while the second term computes the residual between the vector and the cluster centers. The final outputs are obtained by performing intra-normalization and L2 normalization. Discriminative representations emerge because the entire network is trained in an end-to-end manner for speech emotion recognition.

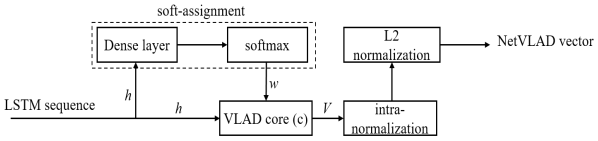


Figure 2: *The workflow of the NetVLAD layer.*

2.2. Label smoothing

We utilize label smoothing to alleviate the problem of imbalanced emotional classes. Label smoothing estimates the marginalized effect of label-dropout during training, and prevents the peaked distributions, which regularizes the model to make it more adaptable.

Specifically, for a example x with label y, the log-likelihood $q(k|x) = \delta_{k,y}$, where $\delta_{k,y}$ is Dirac delta, which equals 1 for $k = y$ and 0 otherwise. The cross entropy loss maximizes the log-likehood.

$$q'(k|x) = (1-\alpha)\,\delta_{k,y} + \alpha u(k) \quad (2)$$

where α is a smoothing parameter. It is weighted mixture of the original ground-truth distribution $q(k|x)$ and the fixed distribution $u(k)$. $u(k)$ is $1/C$ for uniform label smoothing, where C is the number of emotional classes. $u(k)$ is prior class distribution for unigram label smoothing.

3. Database and feature sets

3.1. Dataset

We use Interactive Emotional Dyadic Motion Capture (IEMOCAP) [20] to evaluate our proposed methods. This corpus records (approximately a total of 12 hours) over 5 dyadic sessions with 10 subjects. Each interaction is around 5 minutes in length, and is segmented into sentence levels. We consider only the utterances with majority agreement (at least two out of three evaluators gave the same emotion label). Similar to prior

studies [8][21], the following four emotions are included: "angry", "happy", "sad", and "neutral", with "excited" considered as "happy". In total we use 5,531 utterances: 20.0% "angry", 19.6% "sad", 29.6% "happy", and 30.8% "neutral". The experiment protocol is leave-one-speaker-out and the evaluation metric is unweighted accuracy (UA).

3.2. Feature set

The inputs of LSTM layer are short-time frame-level acoustic features. The feature set is based on the INTERSPEECH 2014 Computational Paralinguistics Challenge [22]. We also add the first dimension of the MFCC, the first order derivatives of all the LLDs, as well as the second order derivatives of MFCC 0-14. The resulting 147 LLDs features are extracted by openSMILE [23].

4. Experiments and analysis

4.1. Experiment settings

We build speech emotion recognition systems based on LSTM model. There is one LSTM layer with 64 memory cells in these systems. We use dropout after LSTM layer with the rate 0.5. The maximum training epochs are 50. The batch size is 32. Adadelta optimization algorithm is utilized. In addition, we inject the Gaussian noise with standard deviation 0.01 into the input features for robust modeling. The dimension of each cluster D is 64. For the unigram smoothing, the unigram prior is computed on the training set.

4.2. Speech emotion recognition based NetVLAD

In this paper, we employ NetVLAD to produce more effective feature representations, which is a concatenation of per cluster residuals weighted by their assignment weights. As a result, the model makes full advantage of temporal sequence information, and generates more emotional-oriented representations. The experimental results, illustrated in Figure 3, show the introduction of NetVLAD is beneficial to performance improvement. We find the models with four clusters achieve best performance 62.6%, and too many or too few clusters would decline the performance. Actually, it conforms with the number of emotional classes, which indicates the clusters are corresponding to the aggregation area of emotional classes implicitly. The weight of soft assignments represents the closeness with different emotional classes.

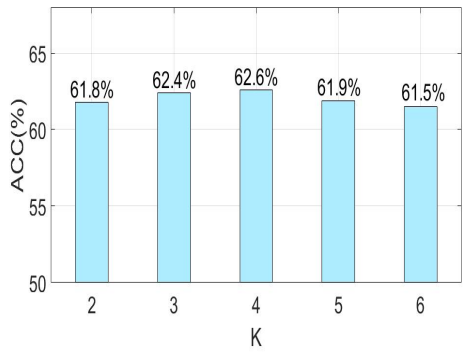


Figure 3: *The performance of the NetVLAD models with different clusters.*

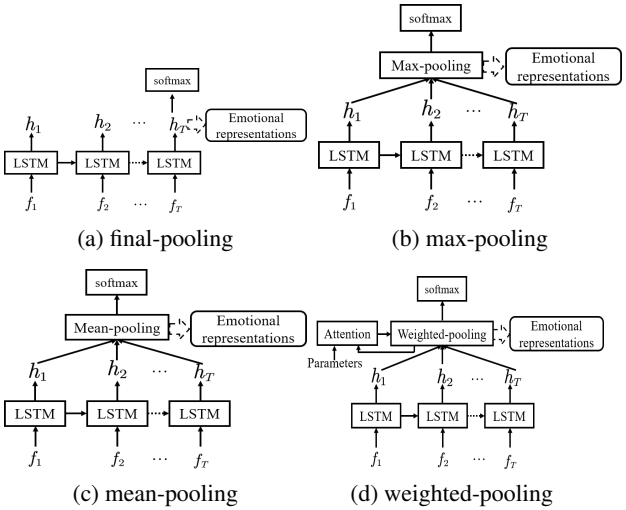


Figure 4: *Different pooling methods.*

We compare the NetVLAD method with other three common adopted sequence-to-label LSTM models. The first is final-pooling shown in the Figure 4(a), which only picks the final state of all hidden states as the emotional representations. The second is max/mean-pooling shown in the Figure 4(b) and 4(c), which calculates the average/maximum of all hidden states as the emotional representations. The third is weighted-pooling shown in the Figure 4(d), which computes a weighted sum of all hidden state as the emotional representations, where the weights are normally determined with an additional attention mechanism.

As shown in Table 1, mean-pooling achieves relatively higher accuracy than final-pooling and max-pooling. Weighted-pooling outperforms other models, probably because more emotionally relevant information is captured by the attention mechanism. The conclusions are similar to the work [7]. However, these pooling methods would lose much temporal information from successive frames inevitably. NetVLAD has the ability to capture emotional content from frame-level feature sequences and achieves better performance than these pooling methods.

Table 1: *The performance of different pooling methods.*

Methods	Accuracy
final-pooling	53.8%
max-pooling	56.8%
mean-pooling	59.6%
weighted-pooling	60.3%
NetVLAD	62.6%

4.3. The effect of label smoothing

Label smoothing encourages the differences between the largest logit and others to become large, which prevents the overfitting and increases the generalization of the models. We explore two types of label smoothing methods based on basic LSTM model with mean-pooling, uniform label smoothing and unigram label smoothing. The experimental results with different α value of

(2) are shown in Figure 5. The performance of the models with label smoothing is better than the model with no label smoothing 59.6%. The performance of two label smoothing methods is comparable when α is 0.1, while unigram label smoothing is superior to uniform label smoothing with larger α. The optimal α is 0.2 and uniform label smoothing achieves the accuracy 62.0%. Therefore, prior class distribution information is helpful to emotional modeling more accurately, further improves the performance.

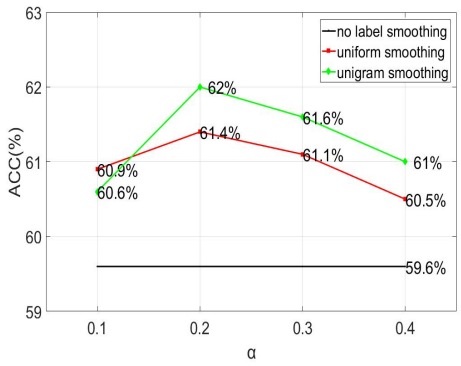
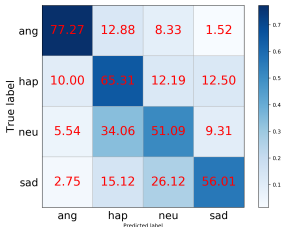
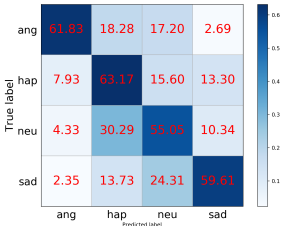


Figure 5: *The performance of different label smoothing methods.*

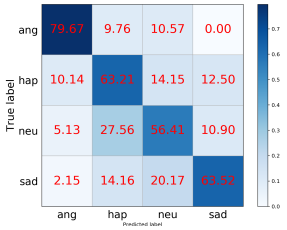
Further, Figure 6 lists the confusion matrixes for no label smoothing 59.6%, uniform label smoothing 61.4% and unigram label smoothing 62.0%. The results reveal no label smoothing has a bias towards angry and happy than neutral and sad. Uniform label smoothing shows more uniform distribution results which reduces the accuracies of angry and happy, and increases the accuracies of neutral and sad. With the help of prior class distribution, the results of unigram label smoothing are also biased, while the accuracies of most classes have been improved compared with no label smoothing.



(a) no label smoothing



(b) uniform label smoothing



(c) unigram label smoothing

Figure 6: *The confusion matrixes of no label smoothing, uniform label smoothing and unigram label smoothing.*

4.4. Comparison

Finally, we combine the NetVLAD and unigram label smoothing strategies together to build the system as shown in Figure 1. We obtain the best performance is 63.5% with four clusters when α is 0.2. Therefore, these two strategies can promote each other and further improve the performance effectively.

We also compare the proposed model with other methods of the literature. Han et al. [4] used deep neural networks to construct utterance-level features, followed by extreme learning machine to obtain 52.1% accuracy. Neumann et al. [24] applied an attentive convolutional neural network with multi-view learning objective function, which achieved 56.1% for speech emotion recognition. Fayek et al. [21] introduced a frame-based formulation to model intra-utterance dynamics with end-to-end deep learning, whose accuracy is 58.1%. The work [8] extracted discriminative embedding features based on a triplet framework with LSTM model, reaching the accuracy 60.4%. Our proposed method achieves better performance than these different pooling methods based on different models, which verifies the effectiveness of generating robust utterance-level representations with NetVLAD and unigram label smoothing to regularize the model.

Table 2: *Performance comparison between our model with other methods.*

Methods	Accuracy
DNN [4]	52.1%
CNN [24]	56.1%
LSTM [21]	58.1%
Triplet framework [8]	60.4%
Proposed model	63.5%

5. Conclusions

In this paper, we utilize NetVLAD to produce more effective feature representations. The models with NetVLAD make full advantage of temporal sequence information, and generate more emotional-oriented representations. The results show the superiority of NetVLAD than general pooling methods. And the optimal number of clusters is four, which is corresponding to the number of emotional classes. It indicates the clusters are corresponding to the aggregation area of emotional classes implicitly. The weight of soft assignments represents the closeness with different emotional classes. In addition, we utilize unigram label smoothing to alleviate the problem of imbalanced emotional classes. The results reveal that unigram label smoothing is better than uniform label smoothing with the help of prior class distribution information. The combination of the NetVLAD and unigram label smoothing further boosts the performance significantly. In the future, we will explore more effective methods to generate robust feature representations for speech emotion recognition.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0822502), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

7. References

[1] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective computing and intelligent interaction*. Springer, 2005, pp. 981–995.

[2] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[3] J. Huang, Y. Li, and J. e. a. Tao, "Effect of dimensional emotion in discrete speech emotion classification," 2017.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

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

[6] L. Chao, J. Tao, and M. e. a. Yang, "Long short term memory recurrent neural network based encoding method for emotion recognition in video," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2752–2756.

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

[8] J. Huang, Y. Li, and J. e. a. Tao, "Speech emotion recognition from variable-length inputs with triplet loss function." in *Interspeech*, 2018, pp. 3673–3677.

[9] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 517–524.

[10] H. Jégou, M. Douze, and C. e. a. Schmid, "Aggregating local descriptors into a compact image representation," in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2010, pp. 3304–3311.

[11] R. Arandjelovic, P. Gronat, and A. e. a. Torii, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[12] W. Xie, A. Nagrani, and J. S. e. a. Chung, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[13] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5150–5154.

[14] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017.

[15] S. Sahu, R. Gupta, and G. e. a. Sivaraman, "Adversarial auto-encoders for speech based emotion recognition," 2017.

[16] Y. Zhang, Y. Zou, J. Peng, D. Luo, and D. Huang, "Discriminative feature learning for speech emotion recognition," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 198–210.

[17] Z. Li, L. He, J. Li, L. Wang, and W.-Q. Zhang, "Towards discriminative representations and unbiased predictions: Class-specific angular softmax for speech emotion recognition," *Proc. Interspeech 2019*, pp. 1696–1700, 2019.

[18] C. Szegedy, V. Vanhoucke, and S. e. a. Ioffe, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[19] G. Pereyra, G. Tucker, and J. e. a. Chorowski, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[20] C. Busso, M. Bulut, and C.-C. e. a. Lee, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[21] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[22] B. Schuller, S. Steidl, and A. e. a. Batliner, "I (special session)******** the interspeech 2014 computational paralinguistics challenge: Cognitive & physical load," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[23] F. Eyben, F. Weninger, and F. e. a. Gross, "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.

[24] M. Neumann and N. T. 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.