Group-level Speech Emotion Recognition Utilising Deep Spectrum Features

Sandra Ottl
sandra.ottl@informatik.uni-augsburg.de
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany

Shahin Amiriparian
shahin.amiriparian@informatik.uni-augsburg.de
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany

Maurice Gerczuk
maurice.gerczuk@informatik.uni-augsburg.de
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany

Vincent Karas
vincent.karas@informatik.uni-augsburg.de
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany

Björn Schuller
schuller@informatik.uni-augsburg.de
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing
Augsburg, Germany

ABSTRACT

The objectives of this challenge paper are twofold: first, we apply a range of neural network based transfer learning approaches to cope with the data scarcity in the field of speech emotion recognition, and second, we fuse the obtained representations and predictions in an early and late fusion strategy to check the complementarity of the applied networks. In particular, we use our Deep Spectrum system to extract deep feature representations from the audio content of the 2020 EmotiW group level emotion prediction challenge data. We evaluate a total of ten ImageNet pre-trained Convolutional Neural Networks, including AlexNet, VGG16, VGG19 and three DenseNet variants as audio feature extractors. We compare their performance to the ComParE feature set used in the challenge baseline, employing simple logistic regression models trained with Stochastic Gradient Descent as classifiers. With the help of late fusion, our approach improves the performance on the test set from 47.88 % to 62.70 % accuracy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Spectral methods**;

KEYWORDS

deep spectrum, pre-trained cnns, emotion recognition, early and late fusion, emotiw

ACM Reference Format:
Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller. 2020. Group-level Speech Emotion Recognition Utilising Deep Spectrum Features. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3382507.3417964

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3417964

1 INTRODUCTION

Emotions play a key role in human interactions and decision-making. Affective computing [25], the interdisciplinary field concerned with automatic detection of emotions and sentiment, has received much attention in recent years due to the variety of research and business opportunities it presents, e. g., for intelligent user interfaces or empathetic digital assistants.

The proliferation of smartphones and other recording devices has led to the availability of large quantities of video material online in which people express emotions. This constitutes "in the wild" data, i. e., data recorded with widely varying conditions (e. g., illumination, occlusion, orientation in the visual, background sounds and reverberation in the audio domain), different sensor characteristics and noisy signals. Such data is more difficult to process than footage of affect gathered under controlled conditions in a lab, and models trained on the latter will usually not generalise well on the former [21]. In addition, the behaviour of subjects in real life situations is naturalistic, with spontaneous displays of affect and complex temporal dynamics. The development of classifiers that can cope with these difficulties is a major challenge in the field of affective computing moving towards real world applications [14].

While many works have studied emotion recognition on the level of individual subjects, in many real world settings people express emotions while interacting with each other. The contextual information contained in those interactions can provide a clue to understanding the emotions of the participants of a conversation. Thus, group level emotion recognition is a promising research field.

Following the creation of the Happy People Images (HAPPEI) database [10], group level emotion recognition was introduced as a subchallenge in the 2016 EmotiW challenge [12]. Variations of the group level emotion recognition problem were featured in the subsequent challenges [9, 11, 13], with the databases transitioning

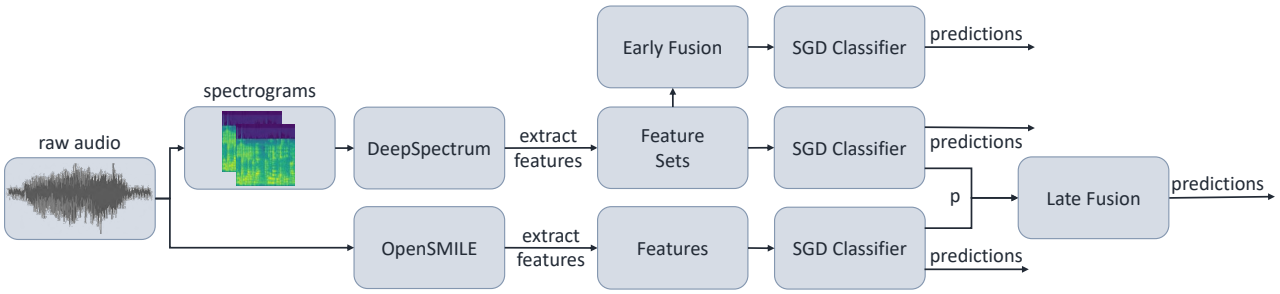


Figure 1: An overview of our proposed transfer learning approach with early and late fusion. First, features are extracted from the raw audio both by openSMILE and by plotting spectrograms that are used as input for our Deep Spectrum system. An SGD classifier is trained on each of the individual feature sets. Moreover, features of different Deep Spectrum nets in varying combinations are fused and later fed into an SGD classifier. Furthermore, the probabilities (p) output by the classifiers are fused.

from collections of images to videos. The current challenge uses the Video level Group AFfect (VGAF) database from [30], which is labelled in terms of group emotion and cohesion.

This paper represents our submission for the eighth Emotion in the Wild "EmotiW 2020" Audio-video based Group Emotion Recognition sub-challenge.

Our approach focuses on speech emotion recognition (SER), where the construction of suitable, robust features is an active area of research [27, 33]. We use our own Deep Spectrum toolkit [1, 4] to learn deep representations from audio spectrograms. Deep Spectrum features have been previously shown to achieve competitive performance on a range of audio recognition tasks [2, 6] while showing resistance to noisy recording conditions [5]. In this process, we employ various popular Convolutional Neural Network (CNN) architectures pre-trained for image recognition. Additionally, we use openSMILE [15] for extracting handcrafted audio features. Finally, we combine deep and handcrafted features using early and late fusion.

The rest of this paper is structured as follows: In Section 2, we describe the dataset. Our method is explained in Section 3. We report our experimental settings in Section 4. The results and their discussion are contained in Section 5. Finally, we give a conclusion and suggestions for future work in Section 6.

2 DATASET

The data given by the challenge organisers contains videos that have been downloaded from YouTube with a creative commons license [30]. The videos feature groups of people speaking and are differing in various aspects such as video quality, number of people and setting. The videos of the train and validation set are labelled in terms of emotional valence with three distinct values (positive, neutral and negative), and it is required to classify the videos of the test set accordingly.

3 APPROACH

A general overview of our CNN-based approach is depicted in Figure 1. First, the mp4 videos are converted into wavs. Then, features are extracted from the audio by openSMILE . More features are acquired by plotting spectrograms and feeding them into our Deep Spectrum system. Next, each of those individual feature sets is used for training a Stochastic Gradient Descent (SGD) classifier. Additionally, features of different combinations of Deep Spectrum nets are fused and afterwards, an SGD classifier is trained on them. Finally, the probabilities of all results are fused, again in varying combinations.

3.1 Feature Extraction

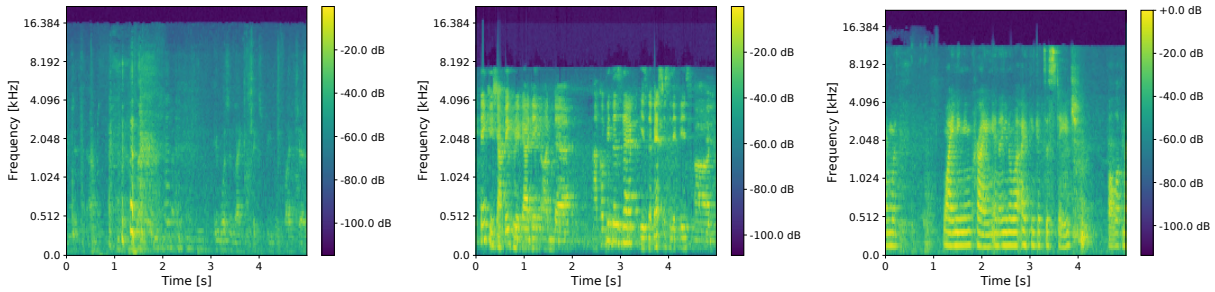
Features are extracted from the videos in two different ways. As described in the first Deep Spectrum part, Mel-spectrograms from the audio instances are used to extract deep representations with different CNN architectures. Secondly, features are extracted with the help of the open-source openSMILE feature extractor [15].

3.1.1 DeepSpectrum. We use the Deep Spectrum system [1, 4] for extraction of deep image-based descriptors from the audio content of the challenge's video recordings. Deep Spectrum is motivated by the efficacy of off-the-shelf CNN descriptors for various visual recognition tasks [24, 29, 35], transferring knowledge from the image domain to audio recognition. The framework has an open-source implementation in the form of a Python toolkit which can be found on GitHub[1]. The first step for this approach is computing Mel-spectrograms from the audio instances. Here, Mel-spectrograms are created by calculating the fast Fourier transform (FFT) on overlapping segments of the audio instances with a width of 2 048 samples and a hop size of 1 024 samples. Reducing the dimensionality of the log-magnitude spectrum with a Mel-filter leads to these Mel-spectrograms. To form the Mel-bands, we use 128 filter banks equally spaced on the Mel-scale:

$$2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \cdot \tag{1}$$

This scale used for displaying frequencies is based on the human perception of frequencies that can disitinguish lower frequencies with a higher resolution [31]. In order to be compatible with image-classification CNNs, the obtained Mel-spectrograms are plotted with the python library Matplotlib [19] and are created with the python library Librosa [23]. Moreover, colour map *viridis* is used for

[1] https://github.com/DeepSpectrum/DeepSpectrum



(a) Positive **(b) Neutral** **(c) Negative**

Figure 2: Example Mel-spectrograms (a – c) computed from audio instances of the train set with ids *12_2*, *10_1* and *18_4*.

visualising the log-magnitudes in the spectrograms. Example Mel-spectrograms from each of the target classes can be seen in Figure 2.

To extract deep representations from these Mel-spectrograms, we use various CNN architectures that will be described in Section 4. All of those networks have been trained on the ImageNet [8] corpus for the task of object classification. The plots of the spectrograms then form the input for these networks and a forward pass is performed for each of them. The neuron activations of a specific layer finally form the audio feature representations for the downstream classification of group level emotion. The choice of feature layer depends on the model architecture but, generally, one of the final layers is used.

3.1.2 openSMILE. Our second approach to extract features is achieved with openSMILE [15]. Here, we use a feature set containing 6 373 features first used in the INTERSPEECH ComParE challenge 2013 [28]. These include voice quality features such as jitter and shimmer, and spectral, ceptral (MFCC), and voicing related low-level descriptors (LLDs). These features are the same ones as used in the challenge baseline, but here, we evaluate them with the same classification models as our Deep Spectrum features for comparison purposes.

3.2 Classification Models

We classify both the Deep Spectrum and openSMILE features by training Stochastic Gradient Descent (SGD) [34] models to minimise either logistic regression or modified Huber losses to which we further add a parameter regularisation term. The models are trained for a total of 1 000 epochs to minimise the combined loss term.

3.3 Fusion

In order to improve on our results with single model approaches, we apply early and late fusion. For early fusion, we fuse our Deep Spectrum and openSMILE descriptors for each the train, validation and test sets by concatenating them along the feature axis. Afterwards, we train the same model as before with those combined features. For late fusion, we observe the class probabilities obtained from each single model on different Deep Spectrum networks and openSMILE features. We perform mean fusion of these probabilities and arrive at the final class predictions by selecting the highest probability.

4 EXPERIMENTAL SETTINGS

We use different ImageNet pre-trained CNN architectures for the extraction of deep feature representations from Mel-spectrograms, as described in Section 3.1.1. First of all, we evaluate the AlexNet [22] and VGG architectures which are composed of standard convolution and maxpooling layers. For VGG, both the 16 and 19 layer variants are tested and for all three networks, the penultimate fully connected layer serves as feature descriptor. Compared to these networks, ResNet50 [17] introduces residual connections which allow for information found in low-level features of early layers to flow upwards through the computation graph and promotes better gradient propagation. DenseNets [18] go a step further by each layer receiving the outputs of all previous layers via feature map concatenation. This allows for the networks to have a more compact architecture which uses fewer feature maps on each layer. For Deep Spectrum , we use three variants with increasing number of dense blocks, denoted by the authors as DenseNet-121, DenseNet-169, and DenseNet-201. Further, we investigate two networks that are built for parameter efficiency, MobileNetV2 [26], and SqueezeNet [20]. Finally, Xception [7] is a CNN which builds on Inception [32] by replacing the inception modules with depthwise-separable convolutions.

All of the Deep Spectrum features are min-max normalised to the range of $[0, 1]$ based on the statistics of the training partition. We further experimented with mean standardisation and no feature scaling at all, but found those to lead to worse results.

For the SGD classifiers, we optimise the choice of training loss – either logistic regression or modified Huber – and the weight and type of parameter regularisation based on validation unweighted average recall (UAR). For regularisation, the squared euclidean norm *l2* or the absolute norm *l1* of model parameters are added to the training loss function with a specific weight factor α. This weight is further optimised on a logarithmic scale from 10^{-6} to 10^{-2} in five steps.

As described in the challenge baseline, accuracy is used as evaluation metric for the task of emotion classification. As the data is unbalanced, we use unweighted average recall (UAR) to find the best model parameters.

Table 1: Comparison of our best performing models with challenge baseline. Performance on validation set (*Val*) and test set (*Test*) is measured in terms of both unweighted average recall (UAR) and accuracy. $^{+}$: significantly better than ComParE on validation ($p < 0.05$). $^{++}$: significantly better than ComParE on validation ($p < 0.01$). As per challenge procedure we evaluated only five of our systems on the test set.

System	Val		Test	
	UAR	Accuracy	UAR	Accuracy
Challenge Baseline [30]	–	50.05	–	47.88
ComParE	53.43	52.48	–	–
DenseNet-121	59.06	56.27	59.89	62.43
Early Fusion of all Deep Spectrum nets	59.17	56.40	–	–
Early Fusion of DenseNet-121 and VGG19 $^{+}$	59.61	56.79	58.54	60.45
Late Fusion of DenseNet-121 and VGG19 $^{+}$	59.75	57.57	59.54	61.91
Late Fusion of DenseNet-121 and VGG19 and ComParE $^{++}$	59.33	57.70	–	–
Late Fusion of all Deep Spectrum nets $^{++}$	59.48	58.09	60.72	**62.70**
Late Fusion of all Deep Spectrum nets and ComParE $^{++}$	60.91	**59.40**	59.90	62.30

Table 2: Comparison of our Deep Spectrum models. *Val*: Results on validation set, measured in both unweighted average recall (UAR) and accuracy. We observe the best performance for DenseNet-121 and VGG19.

Deep Spectrum System	Val	
	UAR	Accuracy
AlexNet	55.18	52.87
DenseNet-121	**59.06**	**56.27**
DenseNet-169	56.19	53.13
DenseNet-201	55.82	53.79
MobileNetV2	53.47	50.52
ResNet50	55.18	53.66
SqueezeNet	56.59	54.96
VGG16	51.78	50.00
VGG19	57.15	54.56
Xception	53.60	51.04

5 RESULTS AND DISCUSSION

We want to focus on the performance of our Deep Spectrum features rather than the classifier. For this purpose, as a first step, we try to compare the success of our system to the features of the baseline by training an SGD classifier on ComParE features we extracted with the help of openSMILE . As depicted in Table 1, this system achieves 53.43 % UAR and 52.48 % accuracy on the validation set, which is slightly higher than the baseline's results on the validation set.

Training an SGD classifier on Deep Spectrum systems employing different CNN architectures achieved UAR and accuracy values shown in Table 2. In terms of UAR, almost all Deep Spectrum descriptors lead to better classification performance than the ComParE acoustic feature set when used as training input for the SGD classifier. These results indicate the suitability of using image-based CNN descriptors for speech emotion recognition in a noisy, in-the-wild, group setting where acoustic parameters are more varied and difficult than in single-speaker speech emotion recognition. The best performing feature representation is extracted from DenseNet-121 arriving at 59.06 % UAR and 56.27 % accuracy on the validation set, and 59.89 % UAR and 62.43 % accuracy on the test set, see Table 1. DenseNets with more dense blocks, DenseNet-169 and DenseNet-201, fall behind. This is contrary to their accuracy on ImageNet [18], where deeper versions achieve higher performance. For speech emotion recognition on the other hand, the features learnt by these networks might be too specific to object recognition. The next best system is accomplished by training an SGD classifier on features coming from VGG19. This results in 57.15 % UAR and 54.56 % accuracy on the validation set. Using VGG16 with 3 less layers than VGG19 arrives only at 51.78 % UAR.

For early fusion, we combined the features of all Deep Spectrum nets and the features of subsets of those. Here, best results were achieved when fusing features obtained with DenseNet-121 and VGG19, resulting in 60.45 % accuracy on the test set, see Table 1.

After training an SGD classifier on all individual Deep Spectrum and ComParE feature sets, we performed mean fusion of their predicted class probabilities. Overall, we can discern that late fusion leads to better results than early fusion on the test set, see Table 1, indicating better generalisation capabilities. Fusing the results for the DenseNet-121 and VGG19 systems achieved 61.91 % accuracy on the test set. Taking ComParE features into this fusion does not improve performance. Fusing the results of all Deep Spectrum systems and ComParE features system arrives at 59.40 % accuracy on the validation set and 62.30 % on the test set. Leaving the ComParE features out of the fusion, the system performs slightly worse on the validation set with an accuracy of 58.09 %. This, however, outperforms the fusion including ComParE features on the test set with 62.70 % accuracy. During validation, we further used a McNemar test for significance, comparing each result against our own ComParE baseline which should be similar to the method employed in the official challenge baseline. We performed the tests at both $p < 0.05$ and $p < 0.01$ and these statistics can be found in Table 1.

6 CONCLUSIONS AND FUTURE WORK

In our contribution to the EmotiW 2020 challenge, we showed the suitability of applying pre-trained image recognition CNNs to audio-based group level emotion recognition. With our Deep Spectrum system, we extracted deep feature representations from the audio content of the challenge dataset using 10 different CNN architectures. Already on their own, most of these representations proved more effective than the challenge baseline's ComParE feature set. Moreover, we could improve the generalisation capabilities and performance of our systems by employing early and late fusion. While we focused on the audio modality, performance improvements could potentially be made to our system by additionally taking the visual content of the videos into account. Furthermore, fusing Deep Spectrum with other unsupervised deep representation learning techniques, such as recurrent autoencoders [3, 16], e. g., auDeep [2] is worth investigating for group level emotion recognition. Finally, we chose to use a relatively simple classification method in our experiments to facilitate fast evaluation of a wide range of Deep Spectrum features. In future work, we want to evaluate combining the Deep Spectrum system with more involved classifiers, such as recurrent neural networks that could help with processing longer audio segments.

ACKNOWLEDGMENTS

This research was partially supported by Zentrales Innovationsprogramm Mittelstand (ZIM) under grant agreement No. 16KN069455 (KIRun), Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 421613952 (ParaStiChaD), BMW AG and the Group on Language, Audio & Music (GLAM) at Imperial College London.

REFERENCES

[1] Shahin Amiriparian. 2019. *Deep Representation Learning Techniques for Audio Signal Processing*. Dissertation. Technische Universität München, München.

[2] S. Amiriparian, N. Cummins, M. Gerczuk, S. Pugachevskiy, S. Ottl, and B. Schuller. 2020. "Are You Playing a Shooter Again?!" Deep Representation Learning for Audio-Based Video Game Genre Recognition. *IEEE Transactions on Games* 12, 2 (2020), 145–154.

[3] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio. In *Proc. of the 2nd Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. IEEE, Munich, Germany, 17–21.

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn W Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *INTERSPEECH*, Vol. 434. ISCA, Stockholm, Sweden, 3512–3516.

[5] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Sergey Pugachevskiy, and Björn Schuller. 2018. Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis. In *Proc. of the 31st International Joint Conference on Neural Networks (IJCNN)*. IEEE, Rio de Janeiro, Brazil, 1–7.

[6] Shahin Amiriparian, Jing Han, Maximilian Schmitt, Alice Baird, Adria Mallol-Ragolta, Manuel Milling, Maurice Gerczuk, and Björn Schuller. 2019. Synchronization in Interpersonal Speech. *Frontiers in Robotics and AI* 6 (2019), 116.

[7] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proc. of the IEEE conference on computer vision and pattern recognition*. IEEE, Honolulu, Hawaii, USA, 1251–1258.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Miami, USA, 248–255.

[9] Abhinav Dhall. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 546–550.

[10] A. Dhall, R. Goecke, and T. Gedeon. 2015. Automatic Group Happiness Intensity Analysis. *IEEE Transactions on Affective Computing* 6, 1 (Jan 2015), 13–26.

[11] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From individual to group-level emotion recognition: EmotiW 5.0. In *Proc. of the 19th ACM International Conference on Multimodal Interaction*, Edward Lank, Eve Hoggan, Sriram Subramanian, Alessandro Vinciarelli, and Stephen A. Brewster (Eds.). Association for Computing Machinery (ACM), United States of America, 524–528. Emotion Recognition in the Wild Challenge (EmotiW).

[12] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. EmotiW 2016: Video and Group-level Emotion Recognition Challenges. In *Proc. of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. ACM, New York, NY, USA, 427–432.

[13] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *Proc. of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 653–656.

[14] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 43.

[15] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462.

[16] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. 2018. auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks. *Journal of Machine Learning Research* 18, 173 (2018), 1–5.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of the IEEE conference on computer vision and pattern recognition*. IEEE, Las Vegas, Nevada, USA, 770–778.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proc. of the IEEE conference on computer vision and pattern recognition*. IEEE, Honolulu, Hawaii, USA, 4700–4708.

[19] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95.

[20] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360 [cs.CV]

[21] Jean Kossaifi, Bjoern W. Schuller, Kam Star, Elnar Hajiyev, Maja Pantic, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, and et al. 2019. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. early access.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Vol. 25. Curran Associates, Inc., Red Hook, NY, USA, 1097–1105.

[23] Brian McFee, Matt McVicar, Oriol Nieto, Stefan Balke, Carl Thome, Dawen Liang, Eric Battenberg, Josh Moore, Rachel Bittner, Ryuichi Yamamoto, Dan Ellis, Fabian-Robert Stoter, Douglas Repetto, Simon Waloschek, CJ Carr, Seth Kranzler, Keunwoo Choi, Petr Viktorin, Joao Felipe Santos, Adrian Holovaty, Waldir Pimenta, and Hojin Lee. 2017. librosa 0.5.0.

[24] Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. 2017. Iris recognition with off-the-shelf CNN features: A deep learning perspective. *IEEE Access* 6 (2017), 18848–18855.

[25] Rosalind W Picard. 2010. Affective computing: from laughter to IEEE. *IEEE Transactions on Affective Computing* 1, 1 (2010), 11–17.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 4510–4520.

[27] Björn Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61 (04 2018), 90–99.

[28] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, Lyon, France, 148–152.

[29] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proc. of the Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Columbus, Ohio, USA, 806–813.

[30] G. Sharma, S. Ghosh, and A. Dhall. 2019. Automatic Group Level Affect and Cohesion Prediction in Videos. In *Proc. of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, Bratislava, Slovakia, 161–167.

[2] https://github.com/auDeep/auDeep

[31] S. S. Stevens, J. Volkmann, and E. B. Newman. 1937. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America* 8, 3 (1937), 185–190.

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proc. of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, San Francisco, California, USA, 4278–4284.

[33] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Lujiazui, Shanghai, China, 5200–5204.

[34] Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proc. of the 21st International Conference on Machine Learning*. ACM, New York, USA, 116.

[35] Yang Zhong, Josephine Sullivan, and Haibo Li. 2016. Face attribute prediction using off-the-shelf CNN features. In *Proc. of the International Conference on Biometrics (ICB)*. IEEE, Halmstad, Sweden, 1–7.