INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Annotation Confidence vs. Training Sample Size: Trade-off Solution for Partially-Continuous Categorical Emotion Recognition

Elena Ryumina, Oxana Verkholyak, Alexey Karpov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia

ryumina_ev@mail.ru, overkholyak@gmail.com, karpov@iias.spb.su

Abstract

Commonly adapted design of emotional corpora includes multiple annotations for the same instance from several annotators. Most of the previous studies assume the ground truth to be an average between all labels or the most frequently used label. Current study shows that this approach may not be optimal for training. By filtering training data according to the level of annotation agreement, it is possible to increase the performance of the system even on unreliable test samples. However, increasing the annotation confidence inevitably leads to a loss of data. Therefore, balancing the trade-off between annotation quality and sample size requires careful investigation. This study presents experimental findings of audio-visual emotion classification on a recently introduced RAMAS dataset, which contains rich categorical partially-continuous annotation for 6 basic emotions, and reveals important conclusions about optimal formulation of ground truth. By applying the proposed approach, it is possible to achieve classification accuracy of UAR=70.51% on the speech utterances with more than 60% agreement, which surpasses previously reported values on this corpus in the literature.

Index Terms: Audio-visual emotion recognition, continuous categorical annotation, annotation confidence, emotional speech resources, computational paralinguistics

1. Introduction

Annotation of emotional data poses a lot of challenges [1]. Different from other machine learning tasks, e.g. object recognition or classification of humans' traits, such as ethnicity, age and gender, where the ground truth can be easily obtained and does not raise any doubts, the perception of emotions varies from person to person, leaving a lot of room for discussion [2].

The annotation process is often approximated with an additive Gaussian noise signal model, where the evaluation result is seen as a linear superposition of the true emotional category and a noise reflecting both human and technical corruptions. Different estimators can be used to estimate the true category of an instance with k different evaluations. Maximum Likelihood Estimator (MLE) is equivalent to the mean value, where all evaluations are assigned the same weight. Evaluator Weighted Estimator (EWE) calculates weights to rule out unreliable evaluators [3]. Such statistical measures as Cohen's and Fleiss' Kappa and Krippendorff's alpha have been extensively used to evaluate the inter-rator agreement for emotionally-colored speech datasets [4]. Additionally, it is possible to assess label uncertainty and annotator idiosyncrasy using hard and soft emotion label annotation [5]. However, most of these approaches are only applicable if the annotators provide labels for exactly same instances.

RAMAS [6] is a uniquely constructed dataset, which has labels of 6 basic emotional categories for some, but not all frames of the recordings. Annotators were free to choose the beginning and the end of emotional expressions. This resulted in multiple overlapping emotional annotation intervals of different lengths for each recording. Due to this rich annotation structure it is difficult to estimate the true emotional category using conventional methods like MLE and EWE. Therefore, this study is the first attempt at identifying strategies for estimating the true emotional category in partially annotated recordings with continuous frame-level labels. We analyze the effects that the quality of partially-continuous annotation has on the performance of emotion classification systems using both video and audio modalities. The contributions include best practices for annotation and evaluation of emotional corpora with continuous and partially-continuous categorical labels, as well as finding the ground truth for annotation with high level of disagreement between annotators. More over, it will be shown that by carefully choosing the training data, it is possible to improve the performance of the classifier even on highly unreliable test samples. Due to these improvements, classification results obtained in this study surpass previously reported in the literature.

2. Proposed Approach

The purpose of the current study is to determine an optimal balance between the quality of annotation and sample size used for training a model. This is achieved by comparing the performance of different models trained on emotional intervals that correspond to different levels of annotation confidence. The experiments include both audio and video modelling in unimodal and bimodal setups. For the fusion of audio and video modalities, we used a weighted fusion of the probabilities for each class [7, 8]. To fuse predictions from 2 models that output probabilities for 6 emotional classes, we generate weight vector with a dimension of 2×6. The weights are generated randomly as a 1000×2×6 matrix using the Dirichlet distribution. We calculate the value of the Unweighted Average Recall (UAR) at each iteration and determine the optimal weight vectors at the maximum UAR.

2.1. Audio modelling

Two systems are adopted for audio modelling: traditional Machine Learning approach and Neural-Network-based approach. Acoustical features are extracted with widely used openSMILE toolkit [9] both on the frame level and on the utterance level. The Low Level Descriptors (LLD) extracted at the frame level include the 88-dimensional eGeMAPS features and 28-dimensional Mel-Frequency Cepstral Coefficients (MFCC), together with their deltas. The utterance-level functionals are ob-

http://dx.doi.org/10.21437/Interspeech.2021-1636

tained by summarizing the LLDs over the whole utterance as described in [10], which gives a 6373-dimensional feature vector. Due to the high number of components, Principle Component Analysis (PCA) is applied to de-correlate these features and reduce the feature dimensionality. With the number of principle components being another parameter to optimize, inline with the previous experiments [11, 12], 300 components provide an optimal performance. The utterance-level features are modelled with two classification methods, Support Vector Machine (SVM) and Logistic Regression (LR), and the frame-level features are used with the Recurrent Neural Network (RNN) Long Short-Term Memory (LSTM) approach.

2.2. Video modelling

Face region detection is performed with OpenCV computer vision library [13]. The face detector is a Single Shot MultiBox Detector (SSD) [14] with a reduced ResNet-10 architecture [15]. We chose the SSD because it has demonstrated its effectiveness in a face detetion task [16, 17]. The detector is trained on images with 300x300 resolution; therefore, the input data is rescaled to match that resolution and then the detected coordinates of face regions are rescaled again to match the original resolution to keep the image quality.

In order to avoid training saturation on same-like consecutive video frames, the frame sequences are pruned by selecting every 5th frame for videos with 50 fps frame rate and every 3-rd frame for videos with 25 fps frame rate. For the first experiment, we extract features from rescaled images of 224x224 pixels using ResNet50 architecture. This model has been trained on VGGFace2 dataset [18], created for recognition of faces, and has been shown efficient for emotion recognition [19]. The features are then passed on to the Random Forest (RF) classifier. For the second experiment, transfer learning technique is used to fine-tune a pretrained model [20]. EfficientNet version B3 [21], which is a type of convolutional neural network, is used as a base model. EfficientNet is one of the pretrained models that was used for extraction of facial features in the framework of EmotioNet Challenge 2020[1] devoted to recognition of in-the-wild emotional expressions. A Fully Connected (FC) layer of size 1024 is added after the penultimate layer of the predefined architecture, and the model is fine-tuned by training all the layers. ReLU activation function and 50% dropout are applied to the last layer. The input data is linearly normalized and rescaled to a resolution of 128x128 to reduce the computational expenses. The model is trained for 10 epochs with Adam optimizer, 0.001 learning rate and a batch size of 32. To obtain the final prediction for a video sequence, probability predictions for every frame in the video are averaged.

3. Experiments

The experiments are conducted on RAMAS [22] dataset with a predefined train/test split. Classification outputs a single label for each emotionally annotated interval. Only the intervals in the recordings that have been annotated are considered for classification. Following are the details about the data and experimental setup.

[1] https://cbcsl.ece.ohio-state.edu/enc-2020/index.html

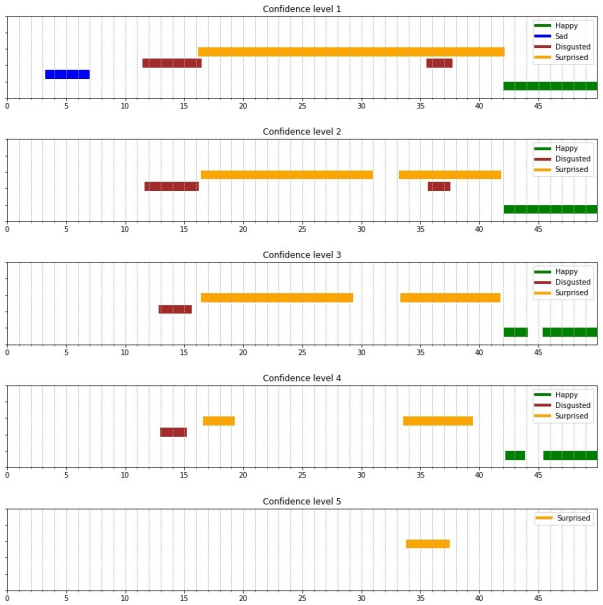

Figure 1: *Annotation of a sample recording from RAMAS according to levels of annotator agreement, from 1 to 5*

3.1. Multimodal data

3.1.1. RAMAS dataset

RAMAS (Russian Acted Multimodal Affective Set) is a Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition [22]. It contains audio-visual recordings of improvised dialogues between two actors. Both male and female actors appear in each scenario. The scenarios are predefined to contain two primary emotions, one per each actor. A total of 6 emotions are considered (Angry, Happy, Surprised, Sad, Disgusted, Scared). Each emotion appears in 4-5 different scenarios. The themes range from daily conversations between friends and coworkers to more elaborate storylines. Actors are given general instructions about what they should discuss in each scenario, however, there are no written scripts. Since actors can improvise within the dialogues, the speech can be considered extemporaneous. In total, there are 6.6 hours, 564 videos and 13 different scenarios played by 10 actors (5 male and 5 female) aged 18-28 years old.

3.1.2. Annotation confidence levels

Each video was labeled by at least 5 different annotators who scored average or higher in an emotional intelligence test [22]. The annotation was performed continuously, i.e. by watching the video and marking the emotions in real time. Since annotators could mark different intervals with different emotion tags, an overlap of emotion labels is common throughout the database. Authors of the database report moderate inter-rater agreement (mean Krippendorff's alpha=0.44). To analyze the ground truth annotation, we compare 5 different levels of label confidence, where level 1 corresponds to at least 1 annotator labeling the interval, and level 5 corresponds to all 5 annotators labeling the interval with the same tag. An example annotation for a given video sample is given in Figure 1. This annotation corresponds to a single speaker; every video has two annotations, one for each speaker, separately.

Table 1: *Number of samples and Krippendorff's alpha according to the level of annotation agreement for each experimental dataset obtained from RAMAS corpus*

Level of agreement	1	2	3	4	5
# Train samples	2448	1632	1462	1369	1130
# Test samples	286	222	236	254	243
Krippendorff's α	0.73	0.81	0.85	0.89	0.94

As can be seen from the figure, there is a compromise between the confidence of annotation and the sample size. The more confident we are in annotation quality, the smaller number of samples is available to train and test the model, and the smaller the size of each sample. The total number of available samples according to the predefined train/test split on each level of annotation agreement, together with the corresponding Krippendorff's alpha value, are show in Table1. To find the balance between the annotation quality and the sample size, we are going to perform a series of experiments to identify the optimal trade-off.

3.1.3. Preprocessing

The neutral emotion was dropped in the initial experiments. We believe it is more useful to perform a preliminary binary classification (emotional/non-emotional speech) followed by the 6-way emotion recognition in case emotional speech is detected in the first stage. This has several advantages. First, the neutral speech tends to last longer then emotional. This distorts the data distribution and makes classification more difficult. Second, 6-way emotion recognition is an easier task than 7-way classification, providing more accurate and reliable results.

After dropping the neutral emotion and removing duplicate entries in the provided labels, we have obtained different emotional intervals for each level of confidence. The number of train and test samples according to each level is depicted in Figure 2. The test set was chosen randomly, making sure that each scenario and emotion are equally represented, and the distribution of labels is comparable to the train set. The train/test split is maintained the same across all experiments. The test set accounted for approximately 10-20% of the whole data, depending on the level of confidence.

The box plots of the lengths of annotated emotional intervals in seconds according to the levels of annotation confidence are shown in Figure 3. The whiskers are set at (1, 99) percentiles. It can be noted that emotional intervals with lower levels of confidence tend to be longer, with more prominent outliers. Intervals with higher level of confidence tend to be shorter, though the difference in lengths is also significant. There are some extremely short instances at each level. For these instances, it was impossible to extract audio features, therefore they were only used in video-based modelling.

3.2. Experimental setup

First, experiments for different levels of annotation confidence were performed independent of each other. We have used both audio and video modalities, as well as their combination, to confirm the experimental results. Since the label distribution in the database is imbalanced, we choose to report Unweighted Average Recall (UAR), also known as weighted accuracy, as a metric for assessing the performance. After individual classification metrics are obtained by testing the systems on the same dataset on which they were trained, we perform the second stage of the experiments by testing the best performing systems trained on each dataset against every other dataset. By doing so we will be able to determine the generalization capability of each model, since systems trained on higher levels of annotation confidence have only been tested on the "good" data, and we want to be able to assess their performance in more general occasions close to real-life scenarios where the data is not perfectly coherent.

3.3. Experimental results

Experimental results for different levels of annotation confidence, performed independently of each other, are shown in Table 2. Thus, we obtained 5 models for different levels of annotation confidence, which were then used for cross-evaluation.

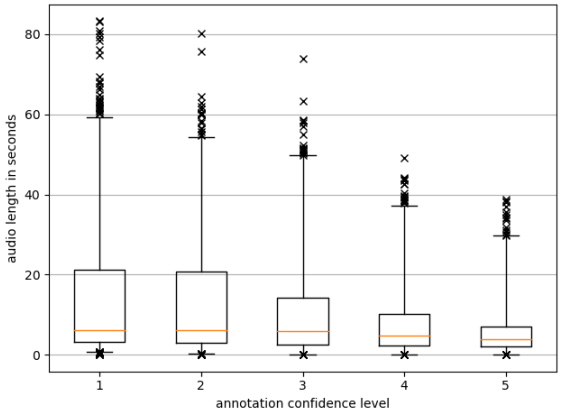

Figure 3: *Emotional interval length (in sec.) according to annotation confidence level*

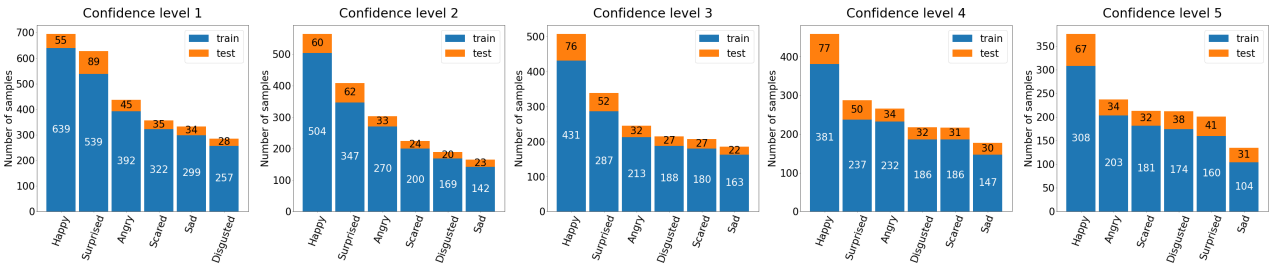

Figure 2: *Number of samples in train and test subsets per each emotional category in the RAMAS dataset according to level of annotation agreement*

As expected, the classification accuracy increases when annotation confidence level increases in all the experiments, allowing the performance to grow from 47.25% to 76.53% in terms of UAR on the same corpus by fusion audio and video modalities. As compared to simple SVM/LR approach, the LSTM neural network provides better classification accuracy on the audio modality due to its high computational power and the ability to model context, however its performance plateaus after the 3rd level of annotation agreement, which can be attributed to a higher number of training parameters. On the other hand, the performance of "lighter" models continues to grow with the annotation confidence level. Different from the audio modality, the performance of video-based systems suffers on the 5-th level of annotation confidence, which can be attributed to reduced sample size, insufficient for optimal training of the models. The loss of training data does not seem to influence audio-based models, whose number of parameters, and therefore the need for data, is significantly lower.

Comparing audio- and video-based systems reveals superiority of the video modality. This can be explained by the nature of annotation process, which assumes watching of the recordings. Since actors tend to use facial expressions not only while speaking, but also while listening to the conversation partner, some of the annotated emotional intervals may not contain speech at all. More over, each microphone dedicated to a single speaker easily picks up speech sounds of the partner, creating a speech overlap and adding a lot of noise to speech samples. However, despite the reasons described above, the fusion of the two modalities gives an increase of/in performance on all 5 levels of annotation confidence.

Experimental results of cross-evaluation of the best performing systems trained on each dataset against every other dataset are shown in Table 3. These results reveal several important findings. First, none of the created models showed significant improvement on the dataset with the lowest level of annotation agreement. This proves yet again that individual ratings are highly subjective and shouldn't be relied upon individually. Second, for optimal performance regardless of quality of test set annotation, training data should be chosen such that it contains emotional intervals where a minimum of 3 annotators agree on the same label. This means that building a classifier on smaller subset of more reliable data (confidence level ≥ 3) allows to increase the performance even on unreliable (confidence level < 3) test data. These results indicate that the common practice used in the emotion classification research (to filter samples on which at least 2 out of 3 annotators agree) may not be optimal. Choosing training data with confidence level = 3 yields an absolute improvement of 4.54% when testing on the dataset with confidence level = 2. Third, the model built on dataset with annotation confidence level = 4 shows overall best performance across all test datasets with different level of annotation agreement.

Table 2: *Classification results (UAR, %) on 5 different models (M) created from RAMAS corpus according to annotation confidence level. Mod. - modality, Feat. - feature type, VGGF - VGGFace2, EN - EfficientNet*

Mod.	Sys.	M 1	M 2	M 3	M 4	M 5
A	SVM	28.87	31.08	38.29	40.43	46.38
A	LSTM	34.38	42.28	46.03	46.28	46.30
V	VGGF	42.96	51.39	57.13	60.19	56.36
V	EN	45.25	53.05	65.26	74.78	70.77
A+V	best	**47.25**	**60.72**	**70.51**	**76.53**	**72.05**

Table 3: *Classification results (UAR, %) of testing 5 best-performing bi-modal (audio LSTM and video-based EfficientNet) models (M) trained on each train subset against every other test subset*

Dataset	Test 1	Test 2	Test 3	Test 4	Test 5
M 1	47.25	54.09	59.10	56.56	57.47
M 2	46.57	60.72	65.57	70.09	76.21
M 3	46.96	65.26	**70.51**	72.14	**77.62**
M 4	45.26	**66.37**	68.15	**76.53**	77.38
M 5	**48.10**	57.76	65.92	73.79	72.05

4. Conclusions

RAMAS is a recently introduced challenging multimodal database that poses a great deal of difficulties for audio- and video-based automatic emotion recognition. The partially-continuous categorical annotation provided in this corpus is a uniquely designed label structure that needs additional considerations since most commonly used estimators of true category are not applicable. Moreover, annotation of this corpus heavily relies on video, making it a leading modality in decision making stage. Audio-based processing shows a subpar performance due to a lot of noise and low correlation with annotation.

Previously, authors of RAMAS database reported the weighted accuracy of 52.5% with stacked bidirectional long short-term memory and decision-level audio and video fusion. One more work reports the weighted accuracy of 65.68% on the emotional intervals with more than 60% of annotator agreement [23]. This corresponds to the 3rd level of annotation confidence as defined in the current study. By training our systems on a higher level of annotation confidence, we were able to reach the performance of 70.51%.

Experiments showed that training a classifier on smaller amount of reliable data boosts the performance even on unreliable test data. In order to achieve an optimal performance on the test data with arbitrary level of annotation confidence, it is imperative to using training data with annotation confidence level ≥ 3. Most of the existing corpora provide annotation from at least 3 annotators, however the common practice is to use samples with confidence level of 2, which may not be optimal. This should be taken into consideration for future emotional corpus design purposes as well. Also, the agreement of 5 or more annotators significantly reduces the sample size and decreases the performance of deep neural networks that are known to be greedy for data. This can be overcome with the increase of data, however it is not always possible. Balancing the trade-off between the quality of annotation and sample size is yet another consideration, which should be taken into account when building emotional classifiers on existing databases.

5. Acknowledgements

This research is partially supported by the Russian Foundation for Basic Research (project No. 20-04-60529) and by the Russian state research (No. 0073-2019-0005).

6. References

[1] M. Kächele, M. Schels, and F. Schwenker, "The influence of annotation, corpus design, and evaluation on the outcome of automatic classification of human emotions," *Frontiers in ICT*, vol. 3, p. 27, 2016.

[2] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human–computer interaction: Comparison and methodological improvements," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.

[3] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 381–385.

[4] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[5] H.-C. Chou and C.-C. Lee, "Every rating matters: joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5886–5890.

[6] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian multimodal corpus of dyadic interaction for affective computing," in *20th International Conference on Speech and Computer SPECOM 2018*, vol. LNAI 11096. Springer, 2018, pp. 501–510.

[7] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *ACM ICMI*, 2015, pp. 459–466.

[8] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, and W. Minker, "An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild," *arXiv preprint arXiv:2010.03692*, 2020.

[9] F. Eyben and B. Schuller, "opensmile:) the munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech 2013*, 2013, pp. 148–152.

[11] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2097–2101. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3160

[12] O. Verkholyak, D. Fedotov, H. Kaya, Y. Zhang, and A. Karpov, "Hierarchical two-level modelling of emotional states in spoken dialog systems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6700–6704.

[13] K. Pulli, A. Baksheev, K. Kornyakov, and V. Eruhimov, "Real-time computer vision with opencv," *Communications of the ACM*, vol. 55, no. 6, pp. 61–69, 2012.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] D. Ryumin, "Automated hand detection method for tasks of gesture recognition in human-machine interfaces," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 20, no. 4, pp. 525–531, 2020.

[17] E. Ryumina, D. Ryumin, D. Ivanko, and A. Karpov, "A novel method for protective face mask detection using convolutional neural networks and image histograms," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIV-2/W1-2021, pp. 177–182, 04 2021.

[18] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018, pp. 67–74.

[19] D. Nguyen, K. Nguyen, S. Sridharan, I. Abbasnejad, D. Dean, and C. Fookes, "Meta transfer learning for facial emotion recognition," in *24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3543–3548.

[20] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," in *Proc. Interspeech 2020*, 2020, pp. 2072–2076. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2666

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[22] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "Ramas: Russian multimodal corpus of dyadic interaction for studying emotion recognition," *PeerJ Preprints*, vol. 6, p. e26688v1, 2018.

[23] D. Fedotov, O. Verkholyak, and A. Karpov, "Contextual continuous recognition of emotions in russian speech using recurrent neural networks," in *Analysis of Verbal Russian Speech (AR3-2019): Proceedings of 8-th Interdisciplinary International Workshop*, 2019, pp. 96–99.