Interspeech 2022
18-22 September 2022, Incheon, Korea

Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus: Multimodal Mask Type Recognition Task

Maxim Markitantov[1], *Elena Ryumina*[1], *Dmitry Ryumin*[1], *Alexey Karpov*[2]

[1]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)
[2]ITMO University, St. Petersburg, Russia

m.markitantov@yandex.ru, ryumina_ev@mail.ru, ryumin.d@iias.spb.su, karpov@iias.spb.su

Abstract

In this paper, we present a new multimodal corpus called Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS), which is designed to analyze voice and facial characteristics of persons wearing various masks, as well as to develop automatic systems for bimodal verification and identification of speakers. In particular, we tackle the multimodal mask type recognition task (6 classes). As a result, audio, visual and multimodal systems were developed, which showed UAR of 54.83%, 72.02% and 82.01%, respectively, on the Test set. These performances are the baseline for the BRAVE-MASKS corpus to compare the follow-up approaches with the proposed systems.

Index Terms: mask type recognition, face masks detection, computational paralinguistics, corpora annotation, data augmentation, machine learning, COVID-19

1. Introduction

Since the end of 2019, all countries have faced the rapid spread of the pandemic caused by COVID-19. The fight against it continues to the present. Many leading scientific groups and global industrial corporations from various scientific fields, such as medicine, biology, computer science, and others, have focused on solving the global problem of reducing the number of infected people around the world [1] and have been conducting research and development of intelligent technologies that will create effective solutions to prevent the spread of COVID-19. Using personal protective equipment or medical masks can reduce the abnormally high rate of spread of respiratory diseases [2, 3]. Systems for automatic analysis of the facial and voice characteristics of a person in a mask can help coping with this issue. Using such a system is not limited only to COVID-19. Masks have been in use in Asia even before the coronavirus [4]. One of the most important steps in developing an automatic recognition system is training data collection. Mask Augsburg Speech Corpus (MASC) [5] and Mask Sorbonne Speech Corpus (MSSC) [6] are speech corpora, which contain audio recordings of speakers with and without medical masks. Moreover, MSSC is a synthetic corpus as it was recorded using a synthetic voice generator held in a block of high-density foam. MAsked FAce (MAFA) [7] and Labeled Faces in the Wild (LFW) [8] are visual corpora that contain images of various faces in masks. However, systems that combine both modalities can greatly improve mask recognition accuracy with respect to systems based on the audio or video information only [9]. Besides, a large amount of multimodal data is required to develop an audio-visual recognition system. Collecting multimodal corpora is not a trivial process. It requires a lot of time, and human resources, participants, as well as special equipment [10]. There exist several bimodal corpora that include some face masks, mostly related to forensics (motorcycle helmets, balaclavas, rubber masks, and hoods with scarfs) [11, 12].

In this paper, we present our Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus. Additionally, we propose a multimodal system for mask type recognition (6 types of masks), based on pre-trained deep neural networks (DNN) that nowadays demonstrate state-of-the-art performance in emotion [13] and speech [14] recognition, speech synthesis [15], image classification [16], object detection [17], as well as they are useful for feature extraction [18].

The layout of the paper is organized as follows. In the next section, we provide a brief background of the methods used in the study. In Section 3, we describe the BRAVE-MASKS corpus and the data collection process. In Section 4, we present recognition system development steps including the proposed pipeline. Experimental results are presented in Section 5, while conclusions are given in Section 6.

2. Background

2.1. Transfer learning

Nowadays, transfer learning [19] is a dominant approach of DNN machine learning when dealing with scarce data in many fields such as computer vision [20], natural language processing [21], and speech recognition [22]. In this study, we use pre-trained audio neural networks (PANNs) [23] that demonstrate state-of-the-art performance in audio pattern recognition.

For the video modality, we use a pre-trained object detector. There are many object detectors, e.g., RetinaNet [24], Yolo [25], R-CNN [26], SPPnet [27], etc. However, the object detectors from the Yolo Family greatly outperform other ones in terms of accuracy and processing speed [28]. In that regard, we used open source codes[1] to train Yolov5 models.

2.2. Data augmentation

Usually a data augmentation procedure allows improving a generalisation ability of models in the face of data scarcity [29, 30, 31]. Specaugment [32] is one of techniques for augmenting audio data. This technique applies a mask on the frequency and time scales. Specaugment simulates a microphone dysfunction in a particular time or signal disappearance in some frequency bands because of an echo.

Mosaic, color space adjustments, and scaling are popular techniques for augmenting video data. Unlike the last two methods, mosaic [33] mixes several random training images. This

[1]https://github.com/ultralytics/yolov5

10.21437/Interspeech.2022-10240

allows combining various classes not presented together in a training set.

3. BRAVE-MASKS corpus

The corpus contains multi-angle images of different person's faces in protective masks of many kinds, as well as audio recordings of continuous Russian speech of people in masks. The multimodal data were recorded using three devices: two Apple iPhone XS Max (left, right) smartphones and an Apple iPad Pro (center) tablet in regular office conditions in front of a heterogeneous background. A Boya BY-M1 microphone was connected to one phone. As shown in Figure 1, three continuous audio-video recordings were made simultaneously. Currently, the corpus contains recordings of 30 native Russian speakers (15 males and 15 females, aged from 19 to 86 y.o., mean age is 40.84, std. dev. is 19.02), both wearing various protective masks and without them. Figure 2 displays age and gender distributions of the informants. The informants performed various tasks and scenarios both without a mask and wearing several different protective masks (disposable medical masks, reusable tissue masks of various colors and prints, medical and special respirators both with and without filters, and protective face shields). In total, different protective masks of 33 types were used (see Figure 3). Then similar protective masks were combined into one class. Thus, we had 6 classes (types of masks): tissue mask (TM), medical mask (MM), FFP2 and FFP3 protective masks (PM), respirator (R), protective face shield (PFS), and no mask (NM). Each informant was recorded in 6 sessions in 3 channels: once without any mask and 5 times wearing 5 different masks. The corpus consists of two parts: bimodal (audio-visual data) and unimodal (video data).

3.1. Bimodal part

The bimodal part contains audio-visual recordings of speech statements. The audio data was sampled at 48 kHz, 16 bit, mono format. Parameters of the video data are equivalent to the unimodal part (see Section 3.2). All the speakers were asked to read sentences from the Russian national standard 50840-95 "Speech transmission over communication paths. Methods for assessing quality, intelligibility and recognition", different for each speaker, and from one phonetically representative text; they also answered some questions and described proposed pictures (e.g., on sport activities, family, kids, food, and countries).

3.2. Unimodal part

Various speaker's head positions and rotations performed during bimodal part recording were not enough to train the object detectors. Therefore we additionally recorded the unimodal part containing only video (without audio) recordings of head rotations (clockwise and counterclockwise) from 8 different points in the room: from a distance of from 0.9 (for audio setup) to 3.2 meters (for video setup) at different angles. Parameters of the video files: resolution of video data is 4K 3840×2160 pixels, frame rate is 60 (for smartphones) and 30 (for tablet) frames per second, color is 24 bits per pixel.

3.3. Corpus annotation

The data of each informant was recorded continuously, so we have split the obtained files into sessions and utterances. Adobe Premiere Pro and Audition were used to synchronize three channels and segment them into sessions and utterances, respectively; after that, the obtained data were further split into Train/Development/Test sets in a speaker-independent way with approximately the same distribution by age and gender classes (see Table 1).

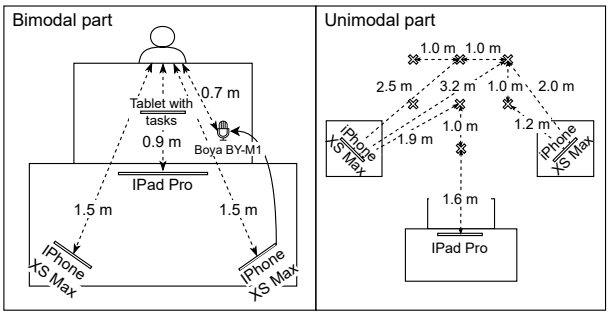

Figure 1: *The BRAVE-MASKS corpus recording setup.*

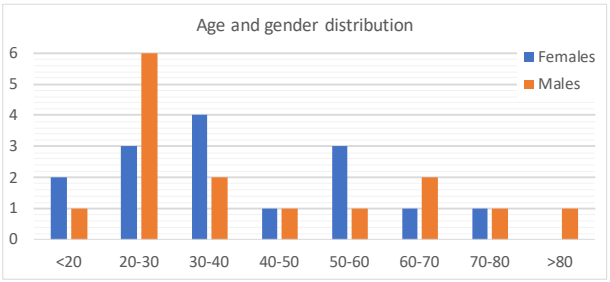

Figure 2: *Informants' age and gender distribution in BRAVE-MASKS.*

For each channel of the bimodal part we obtained 30 speakers × 6 masks × 83 utterances = 14940 video files, in total 21 h 00 min 09 sec. Speakers' utterance length varied from 0.42 to 514.9 sec (for the longest spontaneous narrative).

All recorded video files in the unimodal part with head rotation (30 speakers × 3 channels × 2 rotation scenarios = 180 videos) were cut into fragments for each mask (180 videos × 6 masks = 1080 fragments). A set of one frame per second was extracted from each fragment. Sets of 7800 to 13300 images (the mean is 9350) were extracted from the video recordings of each person in JPG format. Additionally, we have performed a region-of-interest (or mask bounding boxes) annotation. For this, the RetinaFace detector [34] was used. This detector showed its efficiency in our previous research [35]. We found out, that this detector had a lot of false positives (various non-face objects), so we had to manually check annotations for each frame and remove erroneous cases.

Table 1: *Number of instances per mask class in the Train/Development/Test sets for 1 channel. Prot: Protective.*

Class (mask type)	Train	Dev.	Test	∑
No mask (NM)	1328	664	498	2490
Tissue mask (TM)	2490	1328	830	4648
Medical mask (MM)	1079	581	332	1992
Protective mask (PM)	2407	913	996	4316
Respirator (R)	166	166	166	498
Prot. face shield (PFS)	498	664	166	996
∑	7968	3984	2988	14940

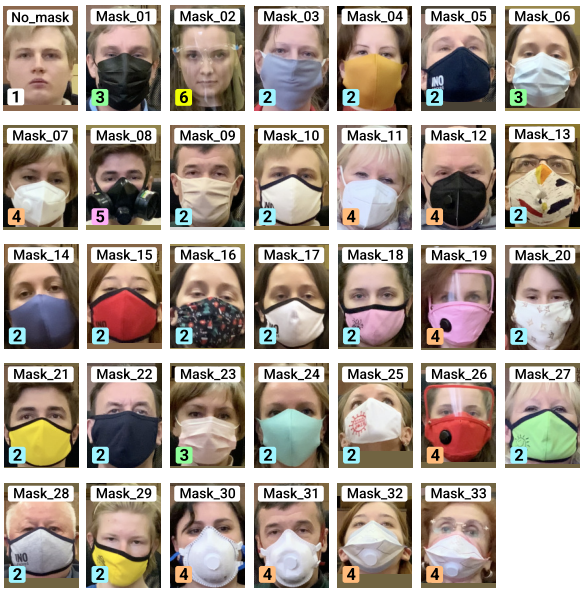

Figure 3: *Samples of informants in various protective masks (mask type in left lower corners)*

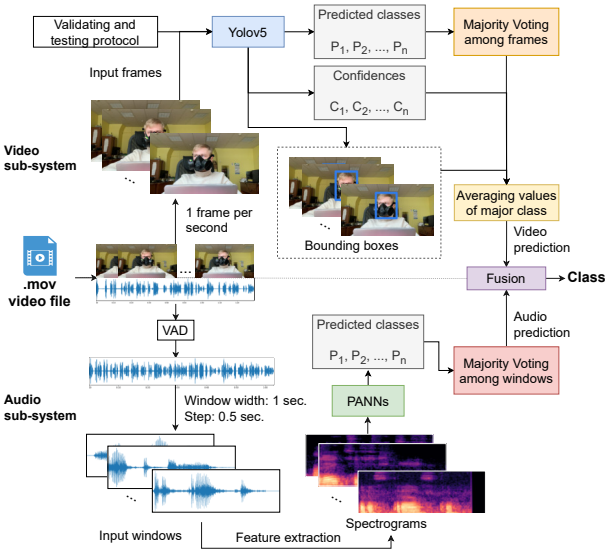

Figure 4: *The pipeline of the proposed multimodal system for mask type recognition. VAD: Voice Activity Detector. PANNs: pre-trained audio neural networks.*

4. Mask Type Recognition Task

The pipeline of the proposed multimodal system for mask type recognition is shown in Figure 4.

All the DNN models were trained on 30 epochs, the mini-batch sizes were 64 (for audio) and 3 (for video). We have used the Adam optimizer with the Cosine Annealing Warm Restarts [36] scheduler. The learning rate ranges from 0.001 to 0.0001. All the models were implemented using the PyTorch software framework [37]. In order to increase the training data size, we used some online data augmentation techniques described above.

We have used a weighted fusion technique [9, 13] to fuse results (hypotheses) of audio and video sub-systems. To do this, a $1000 \times 6 \times 2$ tensor was randomly generated using the Dirichlet distribution, where 1000 is the number of weight matrices, 6 is the number of classes, and 2 is the number of modalities (sub-systems). The weight matrix was chosen on the Development set according to the maximum performance measure. Then, this matrix was applied to the Test set.

4.1. Audio-based Mask Type Recognition

At first we extracted audio data from multimedia files and downsampled them to 16 kHz, after that we used Silero[2] Voice Activity Detector (VAD). Each audio file was cut into some short fragments with the length and step of 1000 ms and 500 ms, respectively. Spectrograms are often used for speech analysis for decades because they provide meaningful information on the voice in the image form [38]. Therefore, we used MelSpectrograms with 64 Mel filterbanks.

From the PANNs [23] we selected 3 CNN models: 6, 10 and 14-layer CNNs. The 6-layer CNN consists of 4 convolutional layers with the kernel size of 5×5. The 10 and 14-layer CNNs consist of 8 and 12 convolutional layers, respectively, with a kernel size of 3×3. The remaining layers are fully-connected. Batch normalization and ReLU non-linearity were applied in between each convolutional layer. Convolutional layers with kernel size equal to 3×3 are grouped by two into one convolutional block. Average pooling was used as the last layer of each convolutional block. After all convolutional blocks a global pooling was applied to summarize feature maps into a fixed-length vector. We used models with pre-trained weights, which are distributed by the authors of the PANNs. The last fully-connected layer was replaced by a new one and obtained models were fine-tuned. In order to fuse the predictions, we used the window-wise majority voting technique.

[2]https://github.com/snakers4/silero-vad

4.2. Video-based Mask Type Recognition

We used images with bounding boxes to train our object detector and score its performance on the Development and Test sets for video modality of both bimodal and unimodal parts. We experimented with several Yolov5 versions and selected Yolov5l because it outperforms shallow versions (Yolov5s, Yolov5m) in terms of accuracy. All Yolov5 versions have 4 modules: Input, Backbone, Neck, and Head. The last three modules are responsible for feature map formation, their transformation (mix and combine) and the prediction of bounding boxes and classes, respectively. Model versions differ in the number of convolutional layers and residual blocks. Yolov5l has 110 and 33, respectively.

Since we used an object detector for the visual mask type recognition method and not a simple classification as in the case of audio, some thresholds were set in the development and test protocol. Face regions were considered at which the thresholds of Intersection over Union (IoU) and confidence were at least 50% and 70%, respectively. If several objects (classes) correspond to this condition, then the one in which the model has the maximum confidence is considered as the correctly predicted object.

Table 2: *Results of the Audio (A) and Video (V) systems on the Development and Test sets (center video channel only). Part: bimodal (B) or unimodal (U) data on which the systems were validated and tested. All performance measures are class-wise unweighted averages (macro) in %. The best results for the audio sub-system are highlighted in bold. *: The result of this sub-system is calculated for 3 channels and is not comparable with the results of other sub-systems in the table.*

				Dev.			Test		
#	**Modality**	**Sub-system / System**	**Part**	**Recall**	**Precision**	**F-score**	**Recall**	**Precision**	**F-score**
1	A	6-layer CNN	B	52.02	60.40	53.68	43.81	62.73	46.49
2	A	10-layer CNN	B	60.36	64.80	60.34	54.65	61.18	54.16
3	A	14-layer CNN	B	**62.07**	64.32	61.74	**54.83**	63.65	56.97
4	V	Yolov5	B	78.11	90.15	75.98	72.02	75.28	73.01
			U*	82.51	87.61	84.53	83.83	88.90	85.96
5	A & V	Fusion of Sub-systems 3 & 4	B	94.32	93.11	93.57	82.01	92.95	85.88

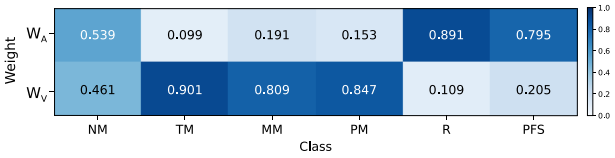

Figure 5: *The weight matrix optimized on the Development set for the final Sub-system 5. W_A: the weights for Sub-system 3. W_V: the weights for Sub-system 4.*

5. Experimental Results

We present baseline experimental results on the Development and Test sets; we did not use these sets to train the models. Moreover, we used only the Development set for neural networks hyper-parameters tuning. Table 2 summarizes results of the proposed systems. We chose the Unweighted Average Recall (UAR) as the baseline performance measure because mask classes in the BRAVE-MASKS corpus are imbalanced.

As can be seen from the results of the audio modality, with increasing model complexity and number of parameters, the performance of the mask type recognition sub-system on the Development and Test sets increases. Thus, the 14-layer CNN (Sub-system 3) showed the best performance. Yolov5 model (Sub-system 4) was trained on the unimodal part and tested on the bimodal and unimodal parts of the BRAVE-MASKS corpus so that researchers could compare the performance of the follow-up approaches with the proposed sub-system without audio modality. Thus, we chose the best sub-systems for the fusion, namely Sub-systems 3 & 4. Since both sub-systems output a class label instead of a probability, we have to transform each prediction of the model into a one-hot vector to weight the results of the two sub-systems. The weight matrix is shown in Figure 5.

It can be seen that the contribution of each sub-system in the final system is approximately the same for the NM class. The Sub-system 5 relies on the results of the Sub-system 4 to predict TM, MM, and PM classes, while the remaining classes are well predicted by the Sub-system 3. This is expectable and well explained by the confusion matrices (see Figure 6). The Sub-system 3 recognizes TM, MM, and PM classes much worse, unlike the Sub-system 4. At the same time, Sub-system 3 confuses MM and TM classes; that means that the acoustic features of these classes are practically the same. The Sub-system 4 could not recognize the R class at all. This is probably due to

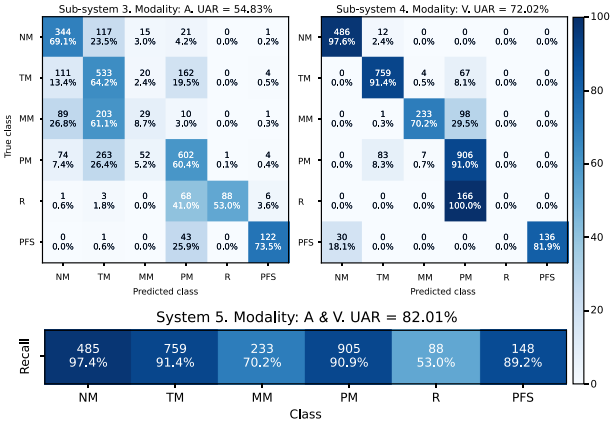

Figure 6: *Confusion matrices of the Sub-system 3, Sub-system 4 and the diagonal elements of the confusion matrix of the System 5 obtained on the Test set. A: audio. V: video.*

the fact that at the training phase for Yolov5 model there were not enough samples of full-face of the R class for a good model generalization. Thus, the combination of the two sub-systems allows compensating weak points of each unimodal sub-system.

6. Conclusions

In this paper, we present a new multimodal corpus called Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) to analyze voice and facial characteristics of persons wearing masks. The detailed description and samples of the BRAVE-MASKS corpus can be found at the web-page[3]. In addition, we propose mask type recognition systems. However, the use of this corpus is not limited only to this task. As a result, audio, visual and multimodal systems were developed for automatic recognition of the mask type (6 classes) of speakers, which showed UAR of 54.83%, 72.02% and 82.01%, respectively, on the Test set.

7. Acknowledgements

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement No. 70-2021-00141.

[3]https://hci.nw.ru/en/pages/brave-masks

8. References

[1] A. Habib, K. M. Anjum, Z. Ashraf *et al.*, "Global epidemiology of covid-19 and the risk of second wave," *Journal of Drug Delivery and Therapeutics*, vol. 11, no. 1, pp. 1–2, 2021.

[2] I. Boškoski, C. Gallo, M. B. Wallace *et al.*, "Covid-19 pandemic and personal protective equipment shortage: protective efficacy comparing masks and scientific methods for respirator reuse," *Gastrointestinal endoscopy*, vol. 92, no. 3, pp. 519–523, 2020.

[3] C. R. MacIntyre and A. A. Chughtai, "Facemasks for the prevention of infection in healthcare and community settings," *Bmj*, vol. 350, pp. 1–12, 2015.

[4] M. Horii, "Why do the japanese wear masks?" *Electronic journal of contemporary Japanese studies*, p. 8, 2014.

[5] M. M. Mohamed, M. A. Nessiem, A. Batliner *et al.*, "Face mask recognition from audio: The masc database and an overview on the mask challenge," *Pattern Recognition*, vol. 122, p. 108361, 2022.

[6] C. Montacié and M.-J. Caraty, "Phonetic, frame clustering and intelligibility analyses for the interspeech 2020 compare challenge." in *Proc. of INTERSPEECH*, 2020, pp. 2062–2066.

[7] S. Ge, J. Li, Q. Ye *et al.*, "Detecting masked faces in the wild with lle-cnns," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2682–2690.

[8] E. Learned-Miller, G. B. Huang, A. RoyChowdhury *et al.*, "Labeled faces in the wild: A survey," in *Advances in face detection and facial image analysis*, 2016, pp. 189–248.

[9] D. Dresvyanskiy, E. Ryumina, H. Kaya *et al.*, "End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild," *Multimodal Technologies and Interaction*, vol. 6, no. 2, 2022.

[10] A. Dvoynikova, M. Markitantov, E. Ryumina *et al.*, "Analytical review of audiovisual systems for determining personal protective equipment on a person's face," *Informatics and Automation*, vol. 20, no. 5, pp. 1116–1152, 2021.

[11] N. Fecher, "The audio-visual face cover corpus: investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear," in *Proc. of INTERSPEECH*, 2012, pp. 2250–2253.

[12] R. Saeidi, T. Niemi, H. Karppelin *et al.*, "Speaker recognition for speech under face cover," in *Proc. of INTERSPEECH*, 2015, pp. 1012–1016.

[13] E. Ryumina, O. Verkholyak, and A. Karpov, "Annotation confidence vs. training sample size: Trade-off solution for partially-continuous categorical emotion recognition," in *Proc. INTERSPEECH*, 2021, pp. 3690–3694.

[14] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[15] Y. Jia, Y. Zhang, R. Weiss *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, p. 11, 2018.

[16] M. Shaha and M. Pawar, "Transfer learning for image classification," in *Proc. of the 2nd international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2018, pp. 656–660.

[17] J. Talukdar, S. Gupta, P. Rajpura *et al.*, "Transfer learning for object detection using state-of-the-art deep neural networks," in *Proc. of the 5th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2018, pp. 78–83.

[18] M. Markitantov, D. Dresvyanskiy, D. Mamontov *et al.*, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," in *Proc. of INTERSPEECH*, 2020, pp. 2072–2076.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[20] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.

[21] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing," *arXiv preprint arXiv:1910.07370*, p. 11, 2019.

[22] C.-X. Qin, D. Qu, and L.-H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–9, 2018.

[23] Q. Kong, Y. Cao, T. Iqbal *et al.*, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[24] T.-Y. Lin, P. Goyal, R. Girshick *et al.*, "Focal loss for dense object detection," in *Proc. of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[25] J. Redmon, S. Divvala, R. Girshick *et al.*, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[26] R. Girshick, J. Donahue, T. Darrell *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] K. He, X. Zhang, S. Ren *et al.*, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[28] Y.-H. Lee and Y. Kim, "Comparison of cnn and yolo for object detection," *Journal of the semiconductor & display technology*, vol. 19, no. 1, pp. 85–92, 2020.

[29] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 309–314.

[30] T. Ko, V. Peddinti, D. Povey *et al.*, "Audio augmentation for speech recognition," in *Proc. of INTERSPEECH*, 2015, pp. 3586–3589.

[31] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[32] D. S. Park, W. Chan, Y. Zhang *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. of INTERSPEECH*, pp. 2613–2617, 2019.

[33] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, p. 17, 2020.

[34] J. Deng, J. Guo, E. Ververas *et al.*, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

[35] E. Ryumina, D. Ryumin, D. Ivanko *et al.*, "A novel method for protective face mask detection using convolutional neural networks and image histograms." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, pp. 177–182, 2021.

[36] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, p. 16, 2016.

[37] A. Paszke, S. Gross, S. Chintala *et al.*, "Automatic differentiation in pytorch," in *Proc. of Autodiff Workshop of Neural Information Processing Systems (NIPS)*, 2017, p. 4.

[38] J. Szep and S. Hariri, "Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion," in *Proc. INTERSPEECH*, 2020, pp. 2087–2091.