INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The MASCFLICHT Corpus: Face Mask Type and Coverage Area Recognition from Speech

Adria Mallol-Ragolta[1,2], *Nils Urbach*[1], *Shuo Liu*[1], *Anton Batliner*[1], *and Björn W. Schuller*[1,2,3]

[1] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Centre for Interdisciplinary Health Research, University of Augsburg, Germany
[3] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

adria.mallol-ragolta@informatik.uni-augsburg.de

Abstract

We present a novel speech dataset for face mask type and coverage area recognition collected with a smartphone. The dataset contains 2 h 27 m 55 s of data from 30 German speakers (15 f, 15 m). The baseline results exploit the functionals of the eGeMAPS feature set, the Mel-spectrogram, and the spectrogram representations of the audio samples. To model the one-dimensional features, we investigate Support Vector Classifiers (SVC) and a neural network classifier. We extract salient information from the two-dimensional representations with Convolutional Neural Network (CNN) based encoders, coupled with a classification block. We use the Unweighted Average Recall (UAR) as the evaluation metric. For the face mask type and the coverage area recognition tasks (3-class problems), the best models on the test partition score a UAR of 49.3 % and 47.8 %, respectively. For the face mask type and coverage area recognition task (5-class problem), the optimal model on the test partition obtains a UAR of 35.0 %.

Index Terms: Face Mask Type Recognition, Face Mask Coverage Area Recognition, Paralinguistics, Health

1. Introduction

Amid the still on-going worldwide pandemic caused by the *Coronavirus Disease 2019* (COVID-19), governments are lifting the obligatoriness of wearing face masks, for instance, in public places or the public transport. Current face mask mandates mainly apply in health facilities, including hospitals, doctors' offices, or pharmacies. Despite the relaxation in the usage of face masks, it is important to remember the effectiveness of this instrument to help control the spread of COVID-19 [1] and reduce the number of COVID-19 deaths [2]. Hence, the development of face mask monitoring tools can, in turn, contribute to the spread control and the mortality rate reduction of airborne diseases, such as COVID-19 [3, 4].

The research on face mask detection exploiting visual and acoustic signals with *Artificial Intelligence* (AI) has been intensified in the last years, motivated by the current pandemic context. In the computer vision literature, transfer learning-based approaches are vastly explored [5, 6, 7, 8, 9]. Related works in this domain also propose hybrid models combining deep and classical machine learning techniques [10]. The release of the *Mask Augsburg Speech Corpus* (MASC) as part of the INTERSPEECH 2020 *Computational Paralinguistics ChallengE* (ComParE) [11] motivated the computer audition research community to start working on this problem [12]. Related works analyse the speech changes produced by wearing a mask [13, 14], use deep neural networks on the spectrogram or the Mel-spectrogram representations of the audio signals [15, 16], and investigate transfer learning-based solutions [17, 18].

The main limitation of the MASC dataset is that it only includes speech samples from participants wearing a surgical face mask. This might be attributed to the pre-pandemic context in which the data was gathered. After the outbreak of COVID-19, we have all learnt that different face mask types offer a different protection level against the virus [19]. Consequently, the research community has redefined the face mask detection problem from speech accordingly. Researchers in [20] consider in their study other types of face masks and include tissue masks, medical masks, FFP2 and FFP3 protective masks, respirators, and protective face shields. However, not only the face mask type is important to control the spread of airborne diseases. Wearing the face mask over the nose and the mouth is also an important aspect [21]. Yet, as this can be more or less annoying, especially for a longer time, some people also tend to wear the mask not 'comme il faut' but, for instance, with the nose uncovered.

In an attempt to reflect this reality, we introduce the *Mask Augsburg Speech Corpus using FFP2 and surgicaL masks with the aIrways Covered Halfway and compleTely* (MASCFLICHT), a novel dataset for face mask type and coverage area recognition from speech collected with a smartphone. This dataset includes audio samples from participants wearing a surgical or an FFP2 face mask only covering the mouth or covering both the mouth and the nose. Herein, we describe the data collection procedure and detail the preliminary experiments conducted to partition the data, so it can be used to benchmark future research. We also provide baseline results exploring the functionals of the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [22] as one-dimensional representations of the audio samples, and Mel-spectrograms and spectrograms as two-dimensional representations. The baseline results tackle the problem from three different perspectives: targeting i) the face mask type recognition, ii) the face mask coverage area recognition, and iii) the face mask type and coverage area recognition.

The rest of the paper is laid out as follows: Section 2 describes the MASCFLICHT Corpus, while Section 3 details the methodology followed. Section 4 compiles and analyses the results obtained, and Section 5 concludes the paper.

2. The MASCFLICHT Corpus

This section introduces the dataset. While Section 2.1 describes the data collection process, Section 2.2 details the pre-processing applied to the raw samples, and the data partitioning procedure.

2.1. Data Collection

The MASCFLICHT Corpus[1] contains 2 h 27 m 55 s of speech samples from 30 German participants (15 f, 15 m) recorded with

[1] https://zenodo.org/record/7985457

10.21437/Interspeech.2023-1438

Table 1: *Summary with the number of F(emale) and M(ale) 1 s-length audio samples available in the MASCFLICHT Corpus per partition and condition. The conditions considered are encoded as: i) NM: No Mask, ii) SP: Surgical Partial — only covering the mouth with a surgical mask, iii) ST: Surgical Total – covering both the mouth and the nose with a surgical mask, iv) FP: FFP2 Partial – only covering the mouth with an FFP2 mask, v) FT: FFP2 Total – covering both the mouth and the nose with an FFP2 mask.*

Condition	**Train**			**Devel**			**Test**			$\sum$
	F	M	$\sum$	F	M	$\sum$	F	M	$\sum$	
NM	571	563	1 134	165	155	320	187	147	334	1 788
SP	548	576	1 124	178	140	318	179	164	343	1 785
ST	549	574	1 123	153	139	292	194	161	355	1 770
FP	562	564	1 126	170	143	313	181	162	343	1 782
FT	572	527	1 099	171	154	325	168	158	326	1 750
$\sum$	2 802	2 804	5 606	837	731	1 568	909	792	1 701	8 875

the microphone embedded in a Xiaomi Mi 10 smartphone. Participants' age ranges from 19 to 55 years old, with a mean age of 25.7 years and a standard deviation of 9.1 years. Prior to the data collection, participants read and signed an *Informed Consent Form* (ICF), which received ethics approval from the university's ethics committee.

Participants' voice was recorded under five different conditions: i) without face mask (NM: No Mask), ii) wearing a surgical mask only with the mouth covered (SP: Surgical Partial), iii) wearing a surgical mask with both the mouth and the nose covered (ST: Surgical Total), iv) wearing an FFP2 mask only with the mouth covered (FP: FFP2 Partial), and v) wearing an FFP2 mask with both the mouth and the nose covered (FT: FFP2 Total). Specifically, we employed the surgical face mask from LyncMed, and the FFP2 face mask from IPOS. To assess whether the aforementioned conditions are detectable from a computational paralinguistics perspective, we collected the data in a controlled environment, i. e., in a quiet room and maintaining the speaker-smartphone distance constant.

We proposed to the participants a free and a guided speech task. The free speech task included describing a picture, while the guided speech task consisted in reading ten sentences. These sentences were selected from the PD1 speech corpus[2] of the *Bavarian Archive for Speech Signals* (BAS), which contains phonetically balanced German sentences. Both tasks were recorded under all the five aforementioned conditions. Additionally, to avoid participants' habituation to the proposed tasks, five different, yet equivalent exercises were created for each task; i. e., we selected five different pictures, and five different sets of sentences. As part of the data collection protocol, we followed a rotation system, so that each tuple of picture and set of sentences was collected under all the five investigated conditions. In other words, we fixed the tasks order, but each participant followed a rotated sequence of the investigated conditions when recording the samples. Each sequence of conditions was repeated every five participants.

2.2. Data Pre-Processing and Partitioning

The unvoiced frames in the recorded speech samples, which might mainly be attributed to speakers' silences during the recordings, do not contain relevant information regarding whether the corresponding speaker wears a face mask. Hence, our first pre-processing steps is the detection and removal of the unvoiced frames. We implement a *Root-Mean-Square* (RMS)-based *Voice Activity Detector* (VAD) with a frame length of 64 ms and a 50 % overlap. We then scale the computed RMS using min-max normalisation. We empirically set a threshold for the RMS of 0.1 to differentiate the voiced from the unvoiced frames. To preserve the naturalness and intelligibility of the processed speech samples, only when the VAD detects at least 10 consecutive unvoiced frames, we remove the corresponding portion of the original signal. Next, we normalise the resulting signal, so its amplitude ranges $\in [-1, 1]$, and discard the last 0.5 s, which contains the acoustic feedback of the smartphone interface when stopping the recording. Analogously to the pre-processing applied in the MASC dataset [11], we finally segment the processed speech samples into chunks of 1 s length. The segmented speech samples are downsampled to 16 kHz, converted to mono, and stored with 16-bit PCM format.

After completing the data pre-processing, we distribute the segmented speech samples into three – train, development, and test – speaker-independent partitions, so that all the samples corresponding to the same speaker are contained in the same partition. In an attempt to balance the recognition difficulty among the partitions, we follow a nested *Leave-One-Speaker-Out Cross-Validation* (LOSO-CV) approach, splitting the samples in the inner loop into five speaker-independent folds, to assess the performance of a preliminary model on each speaker separately. This preliminary model targets the recognition of the five conditions defined in Section 2.1. To this end, we train a linear *Support Vector Classifier* (SVC), prior min-max normalisation of the functionals of the eGeMAPS [22]. We use openSMILE [23], version 2.2.0, to extract the features.

The C parameter in the SVC is defined as the hyperparameter to optimise in this preliminary modelling. At each iteration of the LOSO-CV routine, with the speech samples corresponding to one participant reserved for testing the optimal model, each $C \in [10^{-3}, 10^{-2}, 10^{-1}, 1]$ is used to iteratively train and test a SVC model following a 5-fold cross-validation approach, in accordance to the number of folds in which we distribute the data in the inner loop. We average the *Unweighted Average Recall* (UAR) scores obtained after testing the corresponding model on each one of the five iteratively excluded folds to assign an overall performance to each C parameter. The C value that obtains the highest averaged UAR is defined as the optimal hyperparameter in the current LOSO-CV iteration. The optimal C is used to train the optimal SVC which we finally test on the speech samples corresponding to the initially excluded participant. With this preliminary modelling, analysing the results obtained for each individual participant globally, we achieve an averaged UAR of 32.8 % with a standard deviation of 5.4 %. As the task is tackled as a 5-class classification problem, the chance level in terms of the UAR is 20 %.

The resulting UAR scores per participant computed with the

[2] https://www.bas.uni-muenchen.de/forschung/Bas/BasPD1eng.html

Table 2: *Performance summary of the* ***face mask type recognition*** *models trained in terms of the UAR (%); chance level at 33.3 %. We include the total number of trainable parameters per model, and the optimal model training time.*

Features	Model		UAR [%]		Tr. Time
	Arch.	**# Param.**	**Dev**	**Test**	**(MM:)SS**
eGeMAPS	SVClinear	267	49.1	39.2	34
	SVCrbf	267	49.2	**49.3**	32
	NNC	443	48.5	46.9	18:41
Mel-spec.	CNNscratch	7 363	53.1	43.7	3:12
	RN18scratch	11 179 075	49.8	43.6	5:36
	RN18frozen	2 563	45.6	41.4	5:37
	RN18tuned	11 179 075	54.4	49.2	29:11
Spec.	CNNscratch	7 363	47.6	39.1	1:54
	RN18scratch	11 179 075	51.6	40.5	1:32
	RN18frozen	2 563	45.1	40.7	19:42
	RN18tuned	11 179 075	55.9	49.1	12:33

Table 3: *Performance summary of the* ***face mask coverage area recognition*** *models trained in terms of the UAR (%); chance level at 33.3 %. We include the total number of trainable parameters per model, and the optimal model training time.*

Features	Model		UAR [%]		Tr. Time
	Arch.	**# Param.**	**Dev**	**Test**	**(MM:)SS**
eGeMAPS	SVClinear	267	51.7	45.7	33
	SVCrbf	267	51.2	46.2	35
	NNC	443	53.8	**47.8**	48:43
Mel-spec.	CNNscratch	7 363	47.9	42.6	14:40
	RN18scratch	11 179 075	52.6	40.2	18:53
	RN18frozen	2 563	45.3	38.1	1:51
	RN18tuned	11 179 075	54.1	44.1	11:49
Spec.	CNNscratch	7 363	48.9	42.3	18:35
	RN18scratch	11 179 075	48.9	41.6	1:24
	RN18frozen	2 563	43.4	38.1	22:20
	RN18tuned	11 179 075	52.8	41.2	10:25

preliminary modelling help us distribute the participants among the train, development, and test partitions. Specifically, we select and sort the female participants in descending order in terms of the obtained UAR score. Then, we follow a Round-robin fashion and distribute 60 %, 20 %, and 20 % of the female participants among the train, development, and test partitions, respectively. An analogous procedure is followed to distribute the male participants. With this approach, we not only guarantee that the dataset partitions are gender-balanced, but also that the 'difficulty of recognition' is homogenised across the partitions. To sum up, the MASCFLICHT Corpus contains speech samples from 18 (9 f, 9 m), 6 (3 f, 3 m), and 6 (3 f, 3 m) participants in the train, development, and test partitions, respectively. According to the obtained participants' distribution, we anonymise and randomise the segmented speech samples to populate the train, development, and test partitions. The statistics of the MASCFLICHT Corpus are summarised in Table 1.

3. Methodology

This section describes the methodology followed in this work. Section 3.1 defines the feature representations extracted from the audio samples, Section 3.2 details the implementation and the characteristics of the models assessed, and Section 3.3 summarises the model training routines.

3.1. Feature Extraction

We investigate the performance of the functionals of the eGeMAPS feature set [22], the Mel-spectrograms, and the spectrograms as the representations to extract from the available speech samples, motivated by their successful application in health-related problems found in the literature [24, 25, 26, 27]. We use OPENSMILE [23] to extract the functionals of eGeMAPS, version 2.2.0; they characterise each speech segment with a one-dimensional vector $\in \mathbb{R}^{1\times 88}$. We then compute the Mel-spectrogram representations of the speech samples with the `librosa` library [28], version 0.8.0. As parameters for this calculation, we use 128 mels, a window length of 4 096 samples (256 ms), and a hope size of 64 samples (4 ms). We represent the Mel-spectrograms with a linear frequency scale. Finally, we compute the magnitude of the *Short-Time Fourier Transform* (STFT) of each speech sample using a window length of 4 096 samples (256 ms), and a hope size of 64 samples (4 ms) to obtain its spectrogram representation. The spectrograms are displayed with a logarithmic frequency scale and are also computed using the `librosa` library. The magnitude of both the Mel-spectrogram and the spectrogram representations are converted to dB, min-max normalised, and stored as colour images of 224×224 pixels for further processing.

3.2. Model Description

To model the representations extracted in Section 3.1, we select standard machine learning and deep learning techniques. We investigate a linear kernel-based SVC (SVClinear), and a *Radial Basis Function* (RBF) kernel-based SVC (SVCrbf) to model the functionals of the eGeMAPS. Both models apply min-max normalisation to the input features prior to the model training. We complement the modelling of the one-dimensional data with a neural network classifier (NNC). This network implements one *Fully Connected* (FC) layer – prior batch normalisation of the input features – with 88 neurons at the input and as many neurons at the output as classes we aim our model to recognise.

We explore neural networks composed of an encoder and a classification block to model the two-dimensional representations of the speech samples, i. e., the Mel-spectrograms, and the spectrograms. Specifically, in this work, we compare the performance of two different encoder blocks, which are coupled with the same classification block. The classification block implements the architecture of the aforementioned NNC. The only difference lies in the number of input neurons, which is set to 512 in this case, according to the dimensionality of the embedded representations extracted from the encoder block.

The first encoder (CNNscratch) implements one 2-dimensional convolutional layer with 32 filters, a kernel size of 7×7, and a stride of 1. Following the convolutional layer, we use batch normalisation, and the output is transformed with a *Rectified Linear Unit* (ReLU) function. A 2-dimensional adaptive average pooling layer is implemented at the end to produce a feature map $\in \mathbb{R}^{4\times 4}$ per filter. Finally, we reshape the resulting feature maps, so that the encoder block outputs an embedded representation $\in \mathbb{R}^{1\times 512}$.

The second encoder is based on the ResNet18 architecture [29], but without the last layer. This way, it also produces an embedded representation $\in \mathbb{R}^{1\times 512}$. Coupling this encoder with the classification block motivates the investigation of three different scenarios. The first scenario trains the encoder and the classification blocks from scratch (RN18scratch). The second scenario freezes the corresponding pre-trained weights from the ResNet18 model, and only trains the weights of the classification block (RN18frozen). Finally, the last scenario initialises the

Table 4: *Performance summary of the* ***face mask type and coverage area recognition*** *models trained in terms of the UAR (%); chance level at 20.0 %. We include the total number of trainable parameters per model, and the optimal model training time.*

Features	**Model**		**UAR [%]**		**Tr. Time**
	Arch.	**# Param.**	**Dev**	**Test**	**(MM:)SS**
eGeMAPS	SVClinear	445	31.0	34.1	40
	SVCrbf	445	31.1	33.7	39
	NNC	621	31.9	**35.0**	12:29
Mel-spec.	CNNscratch	8 389	33.3	26.1	30:13
	RN18scratch	11 180 101	31.4	28.9	27:52
	RN18frozen	3 589	30.2	26.4	1:16
	RN18tuned	11 180 101	34.0	33.7	7:28
Spec.	CNNscratch	8 389	32.9	30.5	17:38
	RN18scratch	11 180 101	32.8	32.6	11:34
	RN18frozen	3 589	28.2	26.2	2:46
	RN18tuned	11 180 101	35.1	31.3	7:23

encoder block with the ResNet18 pre-trained weights and trains (fine-tunes) the whole network (RN18tuned).

3.3. Model Training

The parameter to optimise in the SVC-based models is the regularisation parameter C. For each model individually, we conduct grid search among the $C \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]$. The C parameter that scores the highest UAR on the development partition is selected to train the optimal SVC-based model, merging the samples from the train and the development partitions. This optimal model is then assessed on the test data. The SVC-based models are implemented seeding the pseudorandom number generator and using the `scikit-learn` library [30], version 0.24.2.

All neural network-based models are trained under the exact same conditions for a fair comparison. We use the Categorical Cross-Entropy as the loss to minimise, using Adam as the optimiser with a fixed learning rate of 10^{-3}. We select UAR as the evaluation metric and define $\mathcal{L}_{\text{UAR}} = 1 - \text{UAR}$ as the validation error to monitor during training. Network parameters are updated in batches of 64 samples and trained during a maximum of 150 epochs. We implement an early-stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. With this training routine, we aim at determining the optimal number of training epochs that minimise the risk of overfitting. Neural network-based models are implemented with the `PyTorch` library [31], version 1.7.0, seeding the pseudorandom number generator at the models initialisation.

4. Experimental Results

This section presents the baseline results. Model performances are reported in terms of UAR. For comparison purposes, we include the number of trainable parameters per model, and the training time of the optimal model, which merges the samples belonging to the train and the development partitions.

Face Mask Type Recognition. Table 2 summarises the results obtained with the face mask type recognition models. The best UAR on the test partition is obtained with the SVCrbf model exploiting the eGeMAPS, 49.3 %. Competitive results are also achieved when modelling the Mel-Spectrogram and the spectrogram representations with the RN18tuned model, which scores a UAR of 49.2 % and 49.1 % on the test partition, respectively. Furthermore, the RN18tuned model outperforms the other encoder-based architectures. This result indicates the suitability of transfer learning in this problem, which could be attributed to the small sample size of the dataset. Despite the small performance difference among the three top-performing models, there is a huge difference in the number of trainable parameters and in the optimal model training time required, which supports the choice of the SVCrbf model in this case.

Face Mask Coverage Area Recognition. Table 3 compiles the performances achieved by the trained face mask coverage area recognition models. The best UAR on the test partition is obtained with the NNC model exploiting eGeMAPS, 47.8 %. When characterising the speech samples with their Mel-spectrogram representations, the RN18tuned model scores the highest performance on the test partition with a UAR of 44.1 %. Nonetheless, the CNNscratch model achieves the greatest performance with the spectrogram representations on the test partition, 42.3 %. In this case, we observe that the CNNscratch model outperforms the RN18scratch and the RN18frozen models on the test partition. This result could indicate the appropriateness of simple encoders to extract salient information from the two-dimensional representations explored.

Face Mask Type and Coverage Area Recognition. Table 4 collects the model performances scored by the face mask type and coverage area recognition models. The best UAR on the test partition is obtained with the NNC model exploiting the eGeMAPS, 35.0 %. Although the RN18tuned model scores the highest UAR with the Mel-spectrograms, 33.7 %, the RN18scratch model achieves the greatest UAR with the spectrograms, 32.6 %, both on the test partition. Hence, ResNet18-based encoders seem a reasonable choice in this case.

5. Conclusions

We introduced the MASCFLICHT Corpus, a novel dataset for face mask type and coverage area recognition from speech collected using a smartphone. We detailed the pre-processing applied to the raw speech samples and reported the procedure followed to partition the data. In addition to introducing the data, the goal of the paper was to provide baseline results. We described the features extracted from the speech samples and defined the models implemented to train the baselines. The results obtained indicated the suitability of eGeMAPS as the feature representation to extract from the speech samples. When tackling the face mask type recognition problem, the SVCrbf model scored the best UAR, 49.3 %. When addressing the face mask coverage area recognition task, the NNC model obtained the greatest UAR, 47.8 %. Finally, the face mask type and coverage area recognition model that achieved the highest UAR, 35.0%, implemented the NNC architecture. These performances were assessed with the corresponding models on the test partition.

Follow-up studies should include confusion matrices and feature relevance analysis. The baseline results open the door to explore other representations and architectures. Further research could investigate the use of Prototypical Networks with the aim to create a prototypical embedding representative of each class. Future works could also consider investigating personalisation approaches to assess their impact on the presented tasks.

6. Acknowledgements

This project has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS). It has also been supported by the Bavarian Ministry of Science and Arts through the ForDigitHealth project, funded by the Bavarian Research Association on Healthy Use of Digital Technologies and Media.

7. References

[1] I. J. Rao, J. J. Vallon, and M. L. Brandeau, "Effectiveness of Face Masks in Reducing the Spread of COVID-19: A Model-Based Analysis," *Medical Decision Making*, vol. 41, no. 8, pp. 988–1003, 2021.

[2] S. Motallebi, R. C. Cheung, B. Mohit, S. Shahabi, A. A. Tabriz, and S. Moattari, "Modeling COVID-19 mortality across 44 countries: face covering may reduce deaths," *American Journal of Preventive Medicine*, vol. 62, no. 4, pp. 483–491, 2022.

[3] Editorial, "COVID-19 transmission – up in the air," *The Lancet. Respiratory Medicine*, 2020.

[4] J. Riou and C. L. Althaus, "Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020," *Eurosurveillance*, vol. 25, no. 4, 2020, article ID 2000058.

[5] G. Jignesh Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face Mask Detection Using Transfer Learning of InceptionV3," in *Proc. of the Intl. Conf. on Big Data Analytics*. Sonepat, India: Springer, 2020, pp. 81–90.

[6] M. R. Bhuiyan, S. A. Khushbu, and M. S. Islam, "A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3," in *Proc. of the 11th Intl. Conf. on Computing, Communication and Networking Technologies*. Kharagpur, India: IEEE, 2020, article ID 49239.

[7] G. Yang, W. Feng, J. Jin, Q. Lei, X. Li, G. Gui, and W. Wang, "Face Mask Recognition System with YOLOV5 Based on Image Recognition," in *Proc. of the 6th Intl. Conf. on Computer and Communications*. Chengdu, China: IEEE, 2020, pp. 1398–1404.

[8] S. Habib, M. Alsanea, M. Aloraini, H. S. Al-Rawashdeh, M. Islam, and S. Khan, "An Efficient and Effective Deep Learning-Based Model for Real-Time Face Mask Detection," *Sensors*, vol. 22, no. 7, 2022, article ID 2602.

[9] X. Su, M. Gao, J. Ren, Y. Li, M. Dong, and X. Liu, "Face mask detection and classification via deep transfer learning," *Multimedia Tools and Applications*, vol. 81, pp. 4475–4494, 2022.

[10] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, 2021, article ID 108288.

[11] B. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. Deighton MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[12] M. M. Mohamed, M. A. Nessiem, A. Batliner, C. Bergler, S. Hantke, M. Schmitt, A. Baird, A. Mallol-Ragolta, V. Karas, S. Amiriparian, and B. Schuller, "Face Mask Recognition from Audio: The MASC Database and an Overview on the Mask Challenge," *Pattern Recognition*, vol. 122, 2022, 11 pages.

[13] C. Montacié and M.-J. Caraty, "Phonetic, Frame Clustering and Intelligibility Analyses for the INTERSPEECH 2020 ComParE Challenge," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2062–2066.

[14] A. Mallol-Ragolta, S. Liu, and B. Schuller, "The Filtering Effect of Face Masks in their Detection from Speech," in *Proc. of the 43rd Annual Intl. Conf. of the Engineering in Medicine & Biology Society*. Guadalajara, Mexico – Virtual Event: IEEE, 2021, pp. 2079–2082.

[15] J. Szep and S. Hariri, "Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2087–2091.

[16] S. Liu, A. Mallol-Ragolta, T. Yan, K. Qian, E. Parada-Cabaleiro, B. Hu, and B. Schuller, "Capturing Time Dynamics from Speech using Neural Networks for Surgical Masks Detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4291–4302, 2022.

[17] T. Koike, K. Qian, B. Schuller, and Y. Yamamoto, "Learning Higher Representations from Pre-Trained Deep Models with Data Augmentation for the ComParE 2020 Challenge Mask Task," in *Proc. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2047–2051.

[18] 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. of Interspeech*. Shanghai, China: ISCA, 2020, pp. 2072–2076.

[19] C. J. Worby and H.-H. Chang, "Face mask use in the general population and optimal resource allocation during the COVID-19 pandemic," *Nature communications*, vol. 11, 2020, article ID 4049.

[20] M. Markitantov, E. Ryumina, D. Ryumin, and A. Karpov, "Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus: Multimodal Mask Type Recognition Task," in *Proc. of Interspeech*. Incheon, Korea: ISCA, 2022, pp. 1756–1760.

[21] C. J. Hemmer, F. Hufert, S. Siewert, and E. Reisinger, "Protection From COVID-19: The Efficacy of Face Masks," *Deutsches Ärzteblatt International*, vol. 118, no. 5, pp. 59–65, 2021.

[22] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proc. of the 18th Intl. Conf. on Multimedia*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[24] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proc. of the 9th Intl. Workshop on Audio/Visual Emotion Challenge, co-located with the 27th Intl. Conf. on Multimedia*. Niece, France: ACM, 2019, pp. 3–12.

[25] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only," in *Proc. of Interspeech*. Brno, Czechia – Hybrid Event: ISCA, 2021, pp. 3830 – 3834.

[26] A. Mallol-Ragolta, H. Cuesta, E. Gómez, and B. Schuller, "Multi-Type Outer Product-Based Fusion of Respiratory Sounds for Detecting COVID-19," in *Proc. of Interspeech*. Incheon, Korea – Hybrid Event: ISCA, 2022, pp. 2163–2167.

[27] A. Mallol-Ragolta, F. B. Pokorny, K. D. Bartl-Pokorny, A. Semertzidou, and B. Schuller, "Triplet Loss-Based Models for COVID-19 Detection from Vocal Sounds," in *Proc. of the 44th Annual Intl. Conf. of the Engineering in Medicine & Biology Society*. Glasgow, UK: IEEE, 2022, pp. 998–1001.

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proc. of the 14th Python in Science Conference*. Austin, Texas: SciPy, 2015, pp. 18–25.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. of the 33rd Conf. on Neural Information Processing Systems*, vol. 32. Vancouver, Canada: NeurIPS, 2019, 12 pages.