Face mask recognition from audio: the MASC database and an overview on the mask challenge

Mostafa M. Mohamed, Mina A. Nessiem, Anton Batliner, Christian Bergler, Simone Hantke, Maximilian Schmitt, Alice Baird, Adria Mallol-Ragolta, Vincent Karas, Shahin Amiriparian, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Mohamed, Mostafa M., Mina A. Nessiem, Anton Batliner, Christian Bergler, Simone Hantke, Maximilian Schmitt, Alice Baird, et al. 2022. "Face mask recognition from audio: the MASC database and an overview on the mask challenge." *Pattern Recognition* 122: 108361. https://doi.org/10.1016/j.patcog.2021.108361.

Face mask recognition from audio: The MASC database and an overview on the mask challenge

Mostafa M. Mohamed[a,d,*], Mina A. Nessiem[a,d], Anton Batliner[a], Christian Bergler[c], Simone Hantke[e], Maximilian Schmitt[a], Alice Baird[a], Adria Mallol-Ragolta[a], Vincent Karas[a], Shahin Amiriparian[a,e], Björn W. Schuller[a,b,e]

[a] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany*
[b] *GLAM – Group on Language, Audio & Music, Imperial College London, UK*
[c] *Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany*
[d] *AI R&D Team SyncPilot GmbH, Augsburg, Germany*
[e] *audEERING GmbH, Gilching, Germany*

1. Introduction

Wayman [1] defines biometric authentication or *biometrics* as "the automatic identification or identity verification of an individual based on physiological and behavioural characteristics". According to this study, several biometric characteristics are available for biometric system designers to choose from, including but not limited to: "fingerprints, voice, iris, retina, hand, face, handwriting, keystroke, and finger shape" [2]. These can be subdivided into those that require physical contact of some sort for the characteristic to be verified such as fingerprints, handprints, or handwriting, and *contact-less* ones that do not – such as voice, iris, retina, or face.

The sudden outbreak of the COVID-19 pandemic presents a significant challenge to the field of biometrics in two ways. First, the virus stays active on surfaces for a long period of time [3], discouraging users from interacting physically with biometric devices shared between multiple people. Consequently, contact-less biometrics became more crucial in the presence of COVID-19. Second, it also spreads in an airborne fashion [3], prompting health authorities worldwide to urge the general public to wear face masks regularly to reduce the spread of the virus [4]. This everyday use prevents existing facial identification systems from functioning properly, whether they are personal (such as those found in personal computers or mobile phones) or public (found in personnel-restricted areas, such as hospitals or airports). For these reasons,

* Corresponding author.
E-mail address: mostafa.mohamed@student.uni-augsburg.de (M.M. Mohamed).

voice biometrics are among the few suitable contact-less biometrics, as the effects of masks impact them less compared to facial biometrics [5,6]. Furthermore, voice biometrics can be convenient in various contexts for several reasons, e. g., in health care [7]. They are easy to use since any smartphone equipped with a microphone can be utilised [7], and they do not need special training for the users, because there are many scenarios where users already use their smartphones by way of speech communication [7].

Speaker identification and verification systems have been researched for a long time [8]; several benchmark datasets are available in this domain [9,10]. Deep Learning (DL) voice biometrics systems have recently been proposed as well [11]. The performance of speaker identification systems has been found by Saeidi et al. [6] to deteriorate when the conditions under which the systems are trained are mismatched with the conditions under which the systems are evaluated. In the context of masks, this means that it is best to identify mask-wearing speakers with models trained on audio from mask-wearing speakers and non-mask-wearing speakers with models trained without. Consequently, automatically classifying whether a speaker is wearing a mask or not employing voice characteristics can improve voice biometrics systems.

The effect of wearing a mask has been thoroughly studied within other contexts; it has been shown to impact the human-to-human perception of speech, although research results have been contradictory as for the question whether this impact is significant for non-hearing impaired people [12] or not [13]. Llamas et al. [14] conducted a thorough investigation of the acoustic effects of different types of face coverings and whether they affect intelligibility; their findings seem to agree with those of Kawase et al. [15], who conclude that the impact of mask-wearing seems to stem from the loss of the visual information that the brain uses to compensate for degradation in auditory information, if not from a direct effect of the facial coverings on the acoustics themselves. Some studies [16,17] analysed the effects of wearing a mask from an acoustic perspective. They found that the affected frequencies are within the range of 1 – 8 kHz, with the greatest impact being within the range of 2 – 4 kHz. These ranges are related to the ranges required for voice biometrics, namely $<$ 1 kHz and 3 – 4.5 kHz [18]. Audio models that predict whether a speaker is wearing a mask or not can offer insights into the relevant effects of mask-wearing by examining the audio features employed by the models.

Machine Learning (ML), and DL especially, have gained much momentum during the last decade. In the field of image processing, Convolutional Neural Networks (CNNs) [19–21] have been used for image classification. Similarly, CNNs [22], Recurrent Neural Networks (RNNs) [23], and generic audio features [24,25] have also been used for audio classification. Given their capabilities, ML and DL have been used to tackle several issues related to the COVID-19 pandemic and other medicine-related problems like cancer detection [26]. Shuja et al. [27] survey a wide range of datasets concerned with several aspects related to COVID-19, including datasets of medical images, e. g., chest X-ray scans, and audio sounds, e. g., coughing and breathing. There are surveys of ML-based COVID-19 diagnosis by way of speech [28] or by medical images [29]. Examples of speech-based applications can be found in [30,31]. DL has also been successfully applied in biometrics; however, advances in voice biometrics are not as pronounced as those in face biometrics [32,33], which necessitates bridging the gap between the two domains.

In audio processing, *spectrograms* – visual representations of audio signals – are often employed. They allow the modelling of audio problems using computer vision techniques, where CNNs have recently shown substantial advancements [34]. For example, DEEP SPECTRUM utilises pretrained CNNs to extract salient visual features from spectrograms and has proven effective in several tasks like snoring-sound classification [35]. Another technique is *transfer learning*, which uses pretrained CNNs and enhances them further by fine-tuning them to be better suited for specific tasks. This is also employed in [22,36], where a large-scale training of CNNs on audio data was performed from scratch. There is a convergence in several DL-based methodologies between the image and audio domains, and consequently, audio processing has made use of the recent advancements in computer vision using DL.

The INTERSPEECH 2020 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) addressed three new problems within the field of Computational Paralinguistics [37]; one of them was the Mask Sub-Challenge (MSC) where the task was to detect whether a speaker is wearing a surgical mask or not, based only on audio signals [38]. We fully review the approaches and results of MSC, by that

1. Providing a detailed review of face mask detection approaches via voice, which can directly enhance voice biometrics;
2. Bridging the gap between DL and voice biometrics by examining the key strengths of the approaches employed by the top participants of MSC;
3. Giving insight on the effects of wearing masks and their impact on audio signal processing.

The article is structured as follows: We will review MSC in Section 2,– its database, evaluation, and baseline features. Then, we present the approaches and results of the participants in Section 3. Their strengths and weaknesses are discussed in Section 4 as well as their usefulness and limitations for voice biometrics. Furthermore, in Section 5, we showcase an Android-based smartphone app that can be used as a proof of concept to deploy audio-based face mask recognition models; benchmarking of the runtime of the top models is included. Concluding remarks are given in Section 6.

2. The Mask Sub-Challenge (MSC)

The Mask Sub-Challenge (MSC) is one of the challenges held within INTERSPEECH 2020 COMPARE. MSC is concerned with the problem: Given a speech recording of 1 s duration, classify whether the speaker is wearing a mask or not. The Sub-Challenge package included scripts that allowed participants to reproduce the baselines and the evaluation. The audio tracks for the Train, Dev(elopment), and Test sets were included, in addition to the binary labels of the Train and Dev sets. The participants were asked to submit the binary labels of the Test set.

In this section, we describe the background of the MSC: the Mask Augsburg Speech Corpus (MASC), the evaluation metric, and the baseline approaches.

2.1. Mask Augsburg Speech Corpus (MASC)

MASC consists of recordings of 32 German native speakers (16 females, 16 males, age from 20 to 41 years, mean age 25.6 years, standard deviation 4.5 years), wearing a Sentinex Lite surgical mask from manufacturer Lohmann and Rauscher. The recordings took place in a sound-proof audio studio of the centre for media and communication at the University of Passau. The studio is padded with special noise reduction wall elements to reduce noise interference and hence ensure good audio recording quality with less reverberation. The recordings were taken using the C4500 BC large diaphragm condenser microphone from AKG; cf. Fig. 1. The audio was sampled at a rate of 48 kHz with 24 bit resolution.

The participants performed different tasks while wearing a mask as well as without: They read the story “Der Nordwind und die Sonne” (“The Northwind and the Sun”) out loud, answered some questions, repeated prerecorded words after listening to them, read words commonly used in medical operation rooms,

Fig. 1. MASC, recording session.

Table 1
Number of chunks per class in the Train/Dev/Test splits for MSC. Test split distributions were blinded during the ongoing challenge.

#	Train	Dev	Test	Σ
clear	5 353	6 666	5 553	17 572
mask	5 542	7 981	5 459	18 982
Σ	10 895	14 647	11 012	36 554

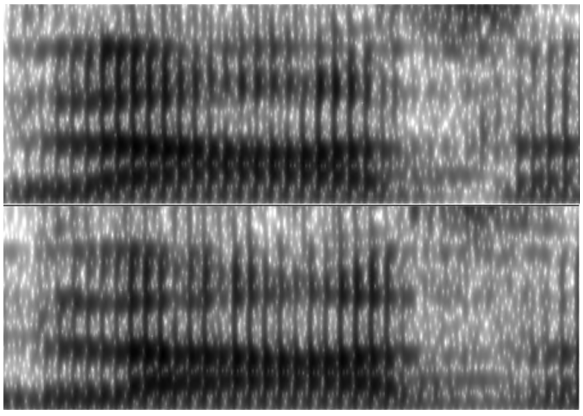


Fig. 2. Spectrograms of the German word "Gallenblase" (*gallbladder*); same male speaker with (below) and without mask (above); shown is 'blase', [bla:s@] in SAMPA transcription; time-aligned signals; displayed are 0.00 - 0.33 sec on the x-axis, 0 - 5000 Hz on the y-axis.

drew a picture and talked about it, and described pictures, e. g., food, sports activities, families, kids, or locations. Both free speech and reading of a defined word list from the medical field are included in the data. The corpus is monolingual (German). Additional meta-data about the identity of the speakers and the speaking tasks were saved for each track.

In order to prepare MASC for MSC, the audio was first downsampled to 16 kHz and converted to mono/16 bit. The data was then partitioned into the usual Train/Dev/Test sets, with 12 speakers in Train and 10 in Dev and Test each. Gender was balanced within each partition, while age differences were disregarded due to the quite homogeneous distribution. The speakers of the three partitions are disjoint; any information about speaker identity was not included in order to be able to classify if a new speaker is wearing a mask or not without any prior information about the speaker. The recordings were chunked consistently into excerpts of 1 s duration without overlap. Chunks without speech were removed from the final data. The total amount of speech in MSC is 10 h 9 min 14 s. Statistics about the sizes of the splits are given in Table 1.

Fig. 2 illustrates differences between masked and clear speech: The same male speaker produces the German word 'Gallenblase' (*gallbladder*) without (above) and with mask (below). Shown is the cut-out segment 'blase', in SAMPA transcription [bla:s@]. The signals are time-aligned. We can see no marked differences; however, especially at the transitions between consonant and vowel – at the beginning: [la:], and at the end: [s@] in schwa (unstressed) position – the masked signal seems to be a bit more blurred. Perceptually, the difference is rather indistinguishable in this setting (see Fig. 1: high-quality microphone, sound-proof room). This will surely change in unfavourable listening conditions, at a distance or with environmental noise.

2.2. Challenge evaluation

In MSC, the performance of the binary classification problem is evaluated using the Unweighted Average Recall (UAR)[1] which is given by:

$$\frac{1}{2}\left(\frac{\mathrm{TP}_{\mathrm{mask}}}{\mathrm{N}_{\mathrm{mask}}}+\frac{\mathrm{TP}_{\mathrm{clear}}}{\mathrm{N}_{\mathrm{clear}}}\right) \tag{1}$$

where $\mathrm{TP}_{\mathrm{mask}}$ is the count of true positives for the *mask* class, $\mathrm{TP}_{\mathrm{clear}}$ is the count of true positives of the *clear* class, $\mathrm{N}_{\mathrm{mask}}$ is the actual number of *mask* examples in the evaluation set, and $\mathrm{N}_{\mathrm{clear}}$ is the actual number of *clear* examples in the evaluation set. In other words, if wearing a mask is the positive class, then $\frac{\mathrm{TP}_{\mathrm{mask}}}{\mathrm{N}_{\mathrm{mask}}}$ is the True Positive Rate (TPR) and $\frac{\mathrm{TP}_{\mathrm{clear}}}{\mathrm{N}_{\mathrm{clear}}}$ is the True Negative Rate (TNR), and accordingly, UAR is the average of TPR and TNR. Consequently, UAR tries to balance both TPR and TNR, similar to other metrics like Area Under ROC Curve (AUC) [40]. UAR facilitates comparison for skewed class distributions; hence it has been used as a standard measure in ComParE since 2009 [41].

There is a correspondence between UAR and typical biometric performance measures like False Match Rate (FMR) and False Non-Match Rate (FNMR), because these are related to the TPR and TNR, respectively. FMR is given by the ratio of false positives with respect to the positive examples, and FNMR is given by the ratio of the false negatives relative to the negative examples [40]. Hence, FMR and FNMR are given by:

$$\mathrm{FMR}=\frac{\mathrm{FP}_{\mathrm{mask}}}{\mathrm{N}_{\mathrm{mask}}}=1[HYPHEN]\mathrm{TPR} \tag{2}$$

$$\mathrm{FNMR}=\frac{\mathrm{FP}_{\mathrm{clear}}}{\mathrm{N}_{\mathrm{clear}}}=1[HYPHEN]\mathrm{TNR} \tag{2}$$

where $\mathrm{FP}_{\mathrm{clear}}$ and $\mathrm{FP}_{\mathrm{mask}}$ are the numbers of falsely classified clear examples and falsely classified mask examples, respectively. Both measures are related to UAR by the equation:

$$\mathrm{UAR}=1-\frac{1}{2}(\mathrm{FMR}+\mathrm{FNMR}) \tag{3}$$

It must be highlighted that measures like FMR and FNMR in biometrics are typically referred to in different contexts, where the classification problem is an acceptance or rejection of transactions or authentication attempts [40]. Nevertheless, we will report the values of these measures for the final results. Moreover, since the labels submitted to MSC are binary labels, only binary classification measures are applicable; measures that evaluate the performance under different thresholds like AUC are not applicable.

2.3. Baseline approaches

In this subsection, we describe different features that are used for the baseline approach; a comprehensive explanation and hyperparameters' exploration are provided by Schuller et al. [38]. The best fusion achieved a UAR of 71.8%.

ComParE Acoustic Feature Set: The official baseline feature set is the same as has been used in previous editions of the ComParE challenges starting from 2013 [42]. It contains 6 373 static features resulting from the computation of 54 functionals (statistics) over 130 low-level descriptor (LLD) contours [42,43]. A full

[1] Note that UAR is sometimes called 'macro-average', see [39].

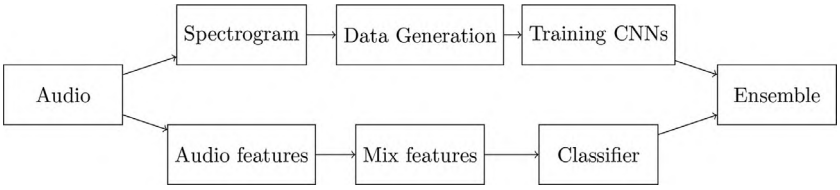


Fig. 3. Illustration of a common generic structure in methodology, which is adapted in some ways by the participants.

description of the feature set, including the LLDs and the functionals, can be found in [24,25]. The LLDs include, but are not limited to, energy, intensity, FFT spectrum, cepstral (MFCC), and psychoacoustic sharpness. The LLDs are obtained on overlapping windows of the audio; then, functionals are applied to reduce the LLDs of the several windows to obtain static features. The functionals include, but are not limited to, moments, extreme values, percentiles, regression coefficients, autoregressive coefficients, and Discrete Cosine Transformation (DCT) coefficients.

Bag-of-Audio-Words (BoAWs): BoAW features were provided as an alternative supra-segmental representation of the LLDs from the ComParE feature set. Instead of applying statistical functionals, a *histogram* of the distribution of the acoustic descriptors is generated. For this, the LLD vector from each frame is *vector quantised* in the first step. A codebook of size N for Vector Quantisation (VQ) is learnt from the training partition by simple *random sampling*. For reproducibility, the toolkit[2] openXBOW [44] is employed, enabling the creation and optimisation of codebooks and BoAW features. Besides ComParE, the BoAW approach has proven its effectiveness in a large variety of audio classification tasks, e. g., acoustic event detection [45].

Deep Spectrum : The feature extraction Deep Spectrum toolkit[3] is applied to obtain high-level deep visual features from the input audio data utilising pretrained CNNs [35]. Mel-spectrograms of the audio are passed through a pretrained ResNet50 [21] (trained on ImageNet), and the activations of the 'avg_pool' layer are extracted, resulting in a 2 048 dimensional feature vector. Deep Spectrum features have been shown to be effective, e. g., for speech processing [46] and audio-based medical applications [47].

auDeep : On the basis of using recurrent Sequence-to-Sequence Autoencoders (S2SAE), the toolkit auDeep[4] constructs unsupervised deep features [48,49]. The learned features explicitly model the inherent sequential nature of the audio signal. Mel-spectrogram representations of the audio are clipped on four power levels below four thresholds X, in order to eliminate the effects of background noise. A distinct S2SAE model is trained for each of these four sets of spectrograms in an unsupervised way, i. e., without any label information. Finally, the learnt encoders' representations are then extracted as 1 024 feature vectors for each audio track, which are concatenated to give fused vectors of 4 096 features.

3. Challenge results and contributions

In this section, we elaborate on the individual approaches of the participants and on the results of fusing their approaches. Many approaches incorporate the baseline features [38] as extra models for their ensembles. Table 2 gives an overview of the performance of all approaches, with some of their highlights.

In Fig. 3, we show an abstract form of a pipeline adapted in one way or another by all participants in the challenge. All techniques either use spectrogram representations or compute some audio features. For spectrogram representations, all approaches make use of the fact that spectrograms transfer the problem to the image domain, where many advanced models can be found. In this scenario, the participants utilise mixtures of several CNNs to classify the given input. ResNet [21] is a common choice, amongst others. Most approaches use *transfer learning* [22,36,50] to adapt pretrained CNNs; thus, their models are trained on much more data in advance. Furthermore, many of the approaches apply techniques for data generation to expand the size of the training data and to overcome the effects of overfitting. These techniques include data augmentations like SpecAugment [51], training a Generative Adversarial Network (GAN) to generate data, or the Mixup technique [52]. Eventually, all the collected features or models are combined and ensembled together to make a final prediction, hence making use of the diversity of the features extracted by each individual model. The ensembling usually relies on simple averaging or majority voting; some approaches use ML models for ensembling such as Support Vector Machines (SVM) [53] for the final predictions.

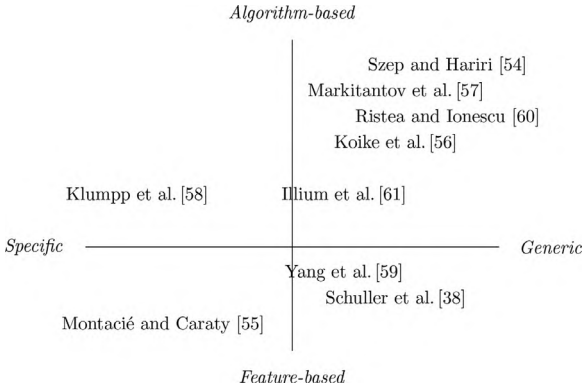


Fig. 4. Comparison of approaches based on how generic they are, and if they are algorithm-based or feature-based. The exact positioning of the approaches is based on the authors' estimate and not on 'objective' criteria. The estimation is based on the definitions given in Section 3, and their rationale can be found at the end of the description of each approach in Section 3.1.

We categorise the approaches according to two main criteria: whether they focus on *features* or *algorithms* and whether the approaches are *generic* or *specific*. Algorithm-based approaches depend primarily on combining several DL techniques in an automated algorithm, while feature-based approaches focus on extracting descriptive (hand-crafted) features with more predictive power.

Generic approaches describe methods that can be adapted to other tasks with only minor changes, while specific approaches deal with methods that have components particularly tailored for the task at hand. A comparison of the approaches regarding these two aspects is shown in Fig. 4. It may not come as a surprise that most of the approaches are localised in the upper right quadrant: Most of the approaches employed are not created for this specific problem but are generic and have been developed for other (types of) problems.

[2] https://github.com/openXBOW/openXBOW
[3] https://github.com/DeepSpectrum/DeepSpectrum
[4] https://github.com/auDeep/auDeep

Table 2
INTERSPEECH 2020 ComParE – MSC Test set results of participants, ordered by performance. The missing ranks either have no submitted paper or they were not accepted by peer-review. Bold text marks the best approach for the specified metric. All metrics are measured in %. The highlights column briefly mentions the key components of the approaches.

Rank	Paper	UAR	FMR	FNMR	Highlights
1	Szep and Hariri [54]	**80.1**	**16.4**	23.4	Multi-band spectrograms, Many CNNs pretrained on ImageNet
3	Montacié and Caraty [55]	77.7	18.5	26.0	Frame and clustered phonetic classes, extracted by SPHINX
4	Koike et al. [56]	77.5	22.7	22.4	CNNs pretrained on AudioSet, Mixup, snapshots during training
6	Markitantov et al. [57]	75.9	**16.4**	31.9	Ensemble of ResNet18 variants, with k-folds and different optimisers
9	Klumpp et al. [58]	75.4	21.8	27.4	RNN for phoneme recognition
11	Yang et al. [59]	75.1	18.5	31.3	Fisher Vectors on ComParE and MFCC, early and late fusions
13	Ristea and Ionescu [60]	74.6	29.2	**21.6**	Cycle-consistent GANs for augmentation
18	Schuller et al. [38]	71.8	16.5	40.0	Deep Spectrum, BoWA, ComParE, auDeep
19	Illium et al. [61]	71.5	34.7	22.3	CNNs with multiple augmentations

3.1. Approaches

We now describe the approaches found in the contributions to the challenge that were accepted for the conference[5]; we first deal with the *algorithm-based* ones, followed by the *feature-based* ones; we first describe the approach chosen by the authors in detail, then we roughly assign to them a position in the two-dimensional space of Fig. 4. Finally, we discuss the key strengths, weaknesses, and findings in Section 4.

3.1.1. Algorithmic-based approaches

Szep and Hariri [54] mainly use several spectrogram features: First, they adopt 3-channel spectrograms with different bandwidths (wide and narrow), with cutoffs at different noise levels: 0 dB and -70 dB. Additionally, they use transfer learning and fine-tuning on three standard image classification CNNs, namely VGG19 [19], DenseNet121 [20], and ResNet101 [21]. These result in 12 combinations of models and features which they ensemble. Furthermore, they merge the Train and Dev sets and train five times using 5-fold cross-validation and ensemble the models resulting from each fold. The use of cross-validation allows the approach to make use of more available data. For the *data generation* part, they utilise simple image augmentation techniques, e. g., rotation (up to 3 degrees) and warping.

The procedure followed in [54] is *generic*, since it is not tailored to the given task and can be used as is for other speech classification tasks. Furthermore, it adapts standard components: mainly state-of-the-art image classification models and features based on several variants of spectrograms, which are generic components that can be applied to any audio data.

Koike et al. [56] make use of several concepts. First, they use a pretrained 14 layers CNN for audio classification, where spectrograms are employed as input method. Second, they use SpecAugment as an audio augmentation mechanism [51], in addition to using the Mixup strategy [52], which mixes input and output examples during training by using a randomised weighted linear combination of different examples. Finally, they adopt an ensemble of several snapshots of the model during training to reduce the effects of overfitting.

The approach in [56] is *generic*, because it consists of several components widely utilised in DL to enhance models and reduce the effect of overfitting. Ensembling with the baseline approaches is not *algorithm-based*; however, the method is still effective without this component.

[5] There are two other approaches found in [62,63] whose authors did not participate in the challenge. We compare few aspects from them to the participants' approaches.

Markitantov et al. [57] submitted five different models to the MSC. These models are all based on two models, ResNet18v1 and ResNet18v2, which are variations of the standard ResNet18 [21]. They make use of four parallel ResNet18s, which are connected to fully-connected layers at the end. The models took as input 64 log-Mel spectrograms and were cross-validated using a variation of k-fold cross-validation with Train and Dev sets shuffled together and split into $k/2$ stratified segments. Their best model is an ensemble of two versions of ResNet18v2, each trained with a different optimisation algorithm.

The approaches introduced in [57] are all *generic* audio-based approaches that depend on variations of the standard ResNet18 model. As such, they can easily be used for other audio tasks without much change.

Ristea and Ionescu [60] use spectrograms as audio representation; however, they employ the real and imaginary components as two separate channels, as opposed to their magnitude as a single channel, which is commonly used. The method trains an ensemble of ResNet models with varying depths, incorporating a novel data augmentation based on cycle-consistent GANs. Eventually, vector representations are generated by the different ResNets and ensemble together using SVM to predict the final classification.

The method is *generic* and can be used for any audio classification task. Furthermore, the method introduces a data augmentation technique, based on training a GAN to generate spectrograms similar to the ones from MASC; this proved to be more effective compared to already existing techniques like SpecAugment [51] – a promising perspective for further experiments.

Illium et al. [61] explore a method that tries using Mel-spectrograms as features representing audio, and then employ some data augmentation technique combined with a CNN in order to solve the classification task at hand. Many augmentation techniques and CNN architectures are explored, and the best combination is used. The augmentations are: speed, loudness, time-shift, random noise, or SpecAugment [51]. From these, time-shifting has proved to be the best suited for the task.

The method provides a *generic* framework for audio classification tasks and does not depend on the task at hand in particular but can be applied for other audio classification tasks.

3.1.2. Phonetic and feature-based approaches

Montacié and Caraty [55] build three different types of models. The first model, Mask Basic System (MBS), uses a variation of the baseline features on the entire audio chunks trained on MASC and an external database Mask Sorbonne Speech Corpus (MSSC). The other two models use phonetic-based features on the level of frames, whole audio chunks, or clusters thereof. Finally, the authors fuse all the models, in addition to the baseline models. The first phonetic-based model operates on a frame basis. The SPHINX

Toolkit [64] is used for each frame to extract phones and the phonetic 'macro-class' (phone/phoneme classes such as front vowels, unvoiced fricatives, or nasals). For each of the 11 (10 + silence) macro-classes, a linear kernel SVM classifier is trained using the ComParE-LLDs of the frames that belong to the macro-class. For the second phonetic-based model, the authors propose a k-means [53] clustering-based model, wherein they cluster the audio frames into k clusters (using ComParE-LLDs as their representation), and for each cluster, they train a linear kernel SVM classifier using the cluster frames. The results are then used on the frame level and the chunk level. The authors perform experiments with the number of clusters ranging from 16 to 1 024.

This approach has some *generic* components like MBS, which uses the baselines features, and partly *specific* features like the phonetic features – partly, because the single phones are language-specific, but phone classes are rather language-generic. Therefore, it can be conceived as a hybrid approach.

Klumpp et al. [58] try to reduce the problem to a phoneme discrimination problem. In other words, wearing masks would affect how phonemes would sound, and hence, the authors use this to classify if the speaker is wearing a mask. First, the method trains a general phoneme recogniser[6] on an external dataset and runs it on MASC; the recogniser classifies 31 phonemes. After that, they distinguish the phonemes of the *mask* examples from the *clear* examples, which results in 62 phonemes, and then trains a second phoneme recogniser on MASC only to recognise the phonemes and if they are spoken with or without mask. This is then used to extract frame features, which are reduced to one vector per sample by computing several functionals. Eventually, a Random Forest classifier [65] is trained to decide whether this vector corresponds to speech with mask or not.

This is a *specific* method that is tailored to the task at hand. An essential part of the training pipeline is based on an assumption of distinguishing between *mask* and *clear*. It might not be straightforward to use this same methodology as is for other audio processing tasks that have nothing to do with speech.

Yang et al. [59] explore Fisher Vector (FV) encoders [66], a widely used computer vision technique that uses a Gaussian Mixture Model (GMM), trained to model the distribution of a set of low-level features. The Fisher Vector encapsulates the first-order and second-order gradients of the features with respect to the GMM model. The authors use two different sets of low level features as training for the Fisher Vector: 13 MFCC features with their first-order and second-order time derivatives, and the 130 ComParE-LLDs – approaches that they term FV-MFCC and FV-ComParE, respectively. Next, they train linear kernel SVMs, while optimising their parameters on the Dev set. Subsequently, a number of experiments are performed, where the different feature sets are fused together, in both early and late fashions.

The method followed by [59] is *generic* as it depends on generic audio features, namely several representations of the ComParE feature set combined with FV, which can be applied for audio processing in general.

3.2. Fusion results

Fig. 5 visualises the fusion of the different approaches. In this case, the best participant's result (80.1% UAR) is defined as 'baseline' at position 1. Gradually, the next best participant's results are iteratively fused by majority vote. In case of a stalemate, the baseline system (position 1) is used, leading to the same UAR when fusing the first two approaches. Since the submitted labels are binary, other fusion techniques cannot be applied. Fig. 5 shows that a fusion of the best five classification systems leads to an absolute improvement of 2.5% UAR compared to the best single approach, resulting in a final and best MASC Test set UAR of 82.6%.

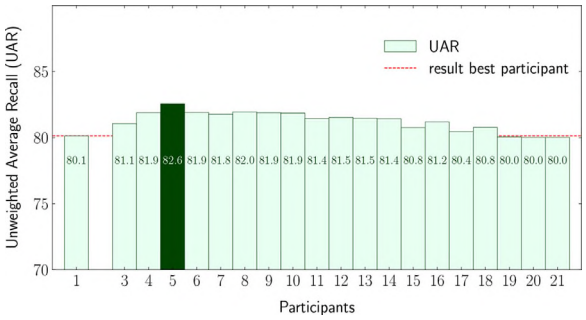


Fig. 5. Fusion results of the best n (1–21) approaches given by each participant following the final MSC ranking shown in Table 2. Fusion calculation only considers classification systems of the 21 participating teams (including not submitted/rejected papers) and not the original baseline system provided by Schuller et al. [38]. Position 2 is empty because in the case of a tie, the winning system is chosen. Best fusion (five systems) is given in dark green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

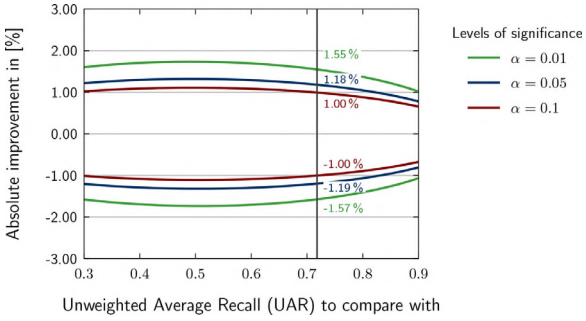


Fig. 6. Two-sided significance test on the MSC Test set with various levels of significance.

In Fig. 5, performance rises till it peaks at five fusions, then it declines more or less slowly to the level of the winning system. In our experience from earlier challenges, fusion does often not pay off when the winning system itself employs several fusion steps. For MSC, the following four systems obviously contribute to modelling variety in the data and thus to the performance.

3.3. Significance test

Fig. 6 visualises a two-sided significance test ([67], chapter 5B) based on the MASC Test set and corresponding baseline system [38]. Various levels of significance (α-values) are utilised in order to calculate an absolute deviation with respect to the Test set, being considered as significantly better or worse than the MSC baseline system. Due to the fact that a two-sided significance test is employed, the α-values must be halved to derive the respective Z-score used to calculate the p-value of a model fulfilling statistical significance for both sides [67]. Consequently, significantly outperforming the MSC baseline system (71.8%, 11 012 Test set samples) at a significance level of $\alpha = 0.01$ requires at least an absolute improvement of 1.55%. Note that Null-Hypothesis-Testing with p-values as criterion has been criticised from its beginning; see the statement of the American Statistical Association in [68,69]. Therefore, we provide this plot with p-values as a service for readers interested in this approach, not as a guideline for deciding between approaches.

[6] Strictly speaking, it is rather a phone and not a phoneme recogniser.

4. Discussion

Based on the characteristics of the approaches detailed in Section 3, we discuss here individual aspects of different approaches.

4.1. Ensemble learning

An essential key ingredient in all approaches is using ensembles, i. e., combining several distinct approaches by employing majority voting, averaging the results, or merging the features and training an extra classifier (SVM in most cases). Ensembling turned out to be successful in all of the approaches. It is typical that ensembling reduces the overall variance of the approach, and consequently gets better results [70]. The top approach in [54] utilises ensembling strongly by merging models trained using 5-fold cross-validation, with 12 models trained for each fold, combining four different spectrogram representations and three different CNN architectures; this results in a large ensemble of 60 models with nearly 1.8 B parameters. Koike et al. [56] and Markitantov et al. [57] utilise ensembles of a particular model with different snapshots during training, either by training several times using different optimisation algorithms or just taking the same model during one training at different steps. In [55,59,60,62], ensembles are used as well. In fact, the approaches by Illium et al. [61] and by [63] are the only approaches that have results worse than the baseline on the Test set, and both of them do not use ensembles. Only [58] managed to get good results without the use of ensembles.[7]

4.2. Transfer learning

An aspect of DL-based approaches is the employment of transfer learning, where pretrained CNNs are fine-tuned, with the pretrained CNNs either being pretrained on image or audio data. This technique has several advantages: it shortens training time, since the networks have already learned to extract salient features before being trained on the task at hand. Another advantage is that it allows the training on large-scale data, which leads to better generalisation by the CNNs pretrained on more diverse data. This is exemplified by Koike et al. [56], who train their CNN using AudioSet [36], which consists of nearly 5 000 hours of audio compared to the total of approximately 10 hours available in MASC.

Szep and Hariri [54] used models that are pretrained on image data; this is similar to the baseline features extracted by Deep Spectrum, which are also based on pretrained image CNNs. Klumpp et al. [58] employed a phoneme recogniser that is pretrained on a larger dataset. Non-DL approaches are based on handcrafted features constructed by human experts for audio processing purposes. This knowledge transfer is utilised implicitly by the feature-based approaches.

All the approaches [54–58,62] that obtained a UAR $\geq$ 75.4% for the Test set are the only ones that used knowledge transfer in some form, either by hand-crafted features or pretrained neural networks.

4.3. Utilising more training data

A common aspect of DL is attempting to increase the size of the training data by employing different techniques. Small data sizes are prone to overfitting; hence, using more data is always recommended for better generalisation of CNN-based models. However, due to the uniqueness and specificity of MASC, the participants could not simply acquire additional data, except for Montacié and Caraty [55], who utilised MSSC. This led them to attempt data generation using different approaches. A common approach for this is employing data augmentation techniques on the input data during training, like SpecAugment [51], or other augmentations like time-shifting and adding Gaussian noise to the input data; this is employed in [61,62]. The most novel idea in this regard is introduced by Ristea and Ionescu [60], who trained a GAN that generates data similar to the training data; this approach surpassed traditional techniques like SpecAugment [51] which masks frequencies and time, as well as image warping. This new technique or other more advanced data augmentation techniques [72] could be utilised in future works. Additionally, Koike et al. [56] used Mixup [52], which generates examples that are linear combinations of pairs of examples from the training data. Furthermore, Szep and Hariri [54] augmented the spectrogram images. Their augmentations consisted of warping (similar to SpecAugment [51]), and minor image rotation (up to 3 degrees). Image augmentations should be carefully applied, such that they do not jeopardise the function of spectrogram visualisations: plotting the time domain against the frequency domain.

4.4. Phonetic features

Montacié and Caraty [55] adopted phonetic features, extracting 45 'phonemes' using the SPHINX toolkit. Klumpp et al. [58] trained an RNN for *phoneme recognition*, i. e., phone recognition[8] on an external larger database. They then used this model to predict 31 phones for a given audio sample, including one model for silence. After that, they distinguished between the phones produced with or without a mask. Phonemes/phones are obviously highly relevant parameters, impacted differently by being filtered through a mask; this is shown by the high performance obtained by Montacié and Caraty [55], who, similarly to [63], break down the results to show the impact on groups of phonemes/phones.

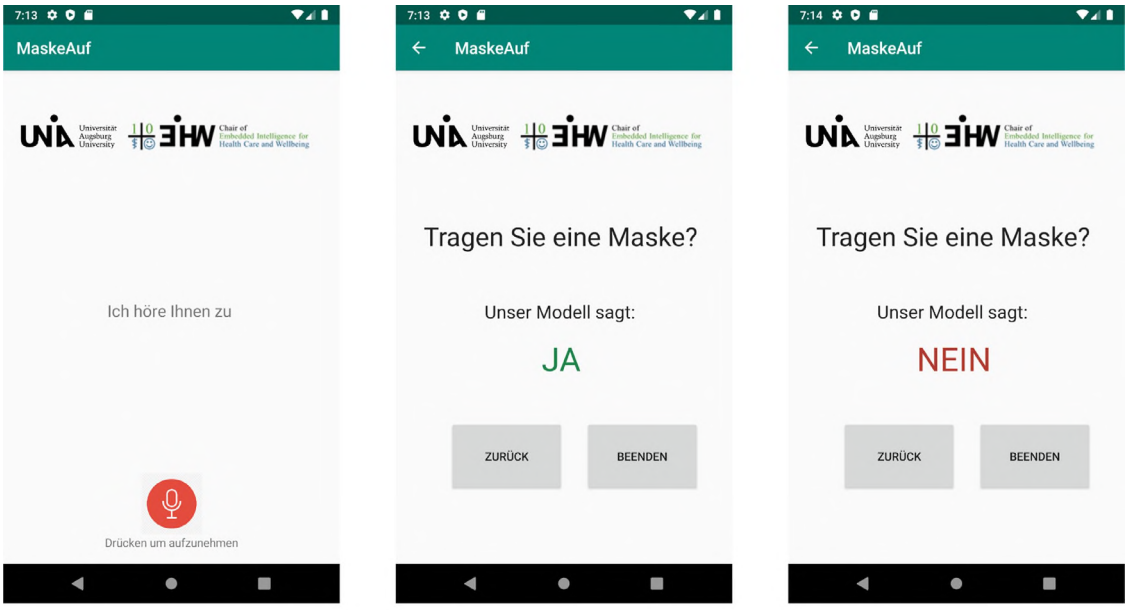
4.5. Interpretation

Some of the approaches were able to extract interpretable information regarding their models. Szep and Hariri [54] extracted a feature map that corresponds in the spectrograms to high neural activation values at the end of the used neural networks. Their analysis shows high activation values around 1 kHz, and the highest activation values are within the frequency range 3 – 5 kHz. The authors suspect that this is why the performance of log-scale spectrograms is degraded compared to linear-scale spectrograms, because the log-scale spectrograms focus more on lower frequency range $<$ 1 kHz than the higher frequencies. This agrees with the attenuation within the range 1 – 8 kHz due to wearing masks, as concluded by [16,17]. The extracted ranges strongly intersect with the ranges relevant for speaker identification, namely $<$ 1 kHz and 3 – 4.5 kHz [18]. Together with earlier studies, these findings suggest that wearing a mask indeed has a general acoustic effect, and a particular effect on tasks in voice biometrics such as speaker identification.

Klumpp et al. [58] conducted an analysis on which phoneme groups were most affected by using masks, the top four groups being *unvoiced plosive, fricatives, approximants* and *vibrants*. We have seen in Fig. 2 that *fricatives* are affected by filtering through the

[7] Note that combining the best results of participants by using late fusion with simple majority voting has often been (slightly) superior to the results of the winning system in former ComParE challenges, see, e. g., [71].

[8] Note that phonemes are underlying, theoretical entities. Although 'phoneme recogniser/recognition' is the common term, we rather should talk about 'phone recognition' – phones are realisations of underlying phonemes and can be modelled acoustically. These phones can be mapped, however, onto phonemes and phoneme classes.



(a) Home screen of the app, where users can record their own voice for analysis.

(b) Results page of the app when it is detected that the user *is* wearing a mask.

(c) Results page of the app when it is detected that the user *is not* wearing a mask.

Fig. 7. Screenshots of the Android smartphone app, depicting the main functionalities.

mask; *approximants* and *vibrants* have 'weak' and variable characteristics and might be prone to the same influences. Montacié and Caraty [55] investigated which phoneme groups are more predictive of using masks or not; they concluded that the top four groups are *diphthongs*, *laterals*, *central vowels*, and *back vowels*. The frequency ranges modelled by [54] and the phoneme classes employed by [58] and [55] cannot be fully mapped onto each other. This might be due to different types of modelling. Nevertheless, overall this proves the relevancy of clustering according to phonetic knowledge.

4.6. Limitations and practical aspects for biometrics

A practical aspect that is not considered in the presented approaches is the run-time of the approaches. The approaches are solely focused on the final performance, which often leads to utilising many models to increase the performance. A complete analysis in this regard is not available; however, we assume that the two approaches with the highest performance are probably the ones that most suffer from the worst run-time. It is not very surprising that these methods have achieved the top performances – yet, real-time processing might be required if it comes to real-life applications. In the case of Szep and Hariri [54], they use three CNN architectures, namely VGG19 (39 M) [19], DenseNet121 (7 M) [20], and ResNet101 (43 M) [21]. For each, there are 20 variants (combinations of four spectrogram variants and five cross-validation folds), which leads to a total of around 1.8 B parameters; we benchmark these architectures in Section 5. A set of CNN models with structures of a low number of parameters [20,73] could probably be utilised instead to reach a better computational efficiency while still maintaining a comparably strong performance; however, a further study would be needed to investigate this. Montacié and Caraty [55], on the other hand, utilised a vast range of features, which included obtaining the baseline features, phonetic features at chunk and frame levels, and clustering them. Running all of these might be computationally intensive at run-time for inference. Furthermore, such an approach is tailored because, in a sense, it performs a form of brute-forcing over many possible features, and it obtains the best ones for the task at hand.

Furthermore, when image processing is applicable, e. g., in multimodal biometrics, it is plausible that it yields better results for mask recognition than audio processing; e. g., Mohan et al. [74] achieve over 98% for classifying whether a person is wearing a mask or not. As a result, this would surpass the models presented in this work. On the other hand, wearing a mask still strongly challenges face biometrics compared to voice biometrics [5,6], even if they are better at classifying masks. This opens the space for applications to switch from using face biometrics to using voice biometrics, or multimodal biometrics combining voice biometrics and other non-face contact-less biometrics; in these two scenarios, the presented models would be of direct aid. In particular, they would be used to automatically select speaker identification or verification models that are fine-tuned for dealing either with mask-wearing speakers or non-mask-wearing speakers, which is expected to deliver the best results [6].

5. Proof of concept demonstration

Current smartphones are pocket-sized computers. They can provide biometric functionalities, replacing dedicated devices. Enabling users to access biometric information on their own smartphone has the additional benefit of limiting the need for physical interaction with shared devices, which could help reduce the spread of, for instance, COVID-19 through contaminated surfaces.

As a proof of concept demonstrator, we have implemented an Android-based smartphone app to deploy the audio-based face mask detection models summarised in this work in real-life scenarios. The app implements a microphone functionality for users to record their own voice (cf. Fig. 7a). The code of the application is available open-source[9]

[9] Upon acceptance, a corresponding GitHub repository link will be added (https://github.com/EIHW/MaskDemoApp).

Table 3
Benchmarks of the 3 architectures by the top participants [54]. Measured on 1 s frames or 12 s tracks (16 overlapping frames). Latency is the time needed to process a request. Throughput is the amount of requests that are processed (in parallel) per second. Latency values have 95% confidence intervals of ±0.04 s/request, regardless if a track or frame is sent.

Architecture	Latency (s/request)				Throughput (requests/s)			
	GPU		CPU		GPU		CPU	
	frame	track	frame	track	frame	track	frame	track
ResNet101	0.09	0.89	0.19	2.53	46.98	4.13	9.28	0.86
DenseNet121	0.09	0.86	0.14	1.81	35.45	4.19	15.70	1.11
VGG19	0.08	0.88	0.27	3.85	42.97	4.03	5.09	0.71

Once the recording is completed, the media file is transferred through network to a dedicated server. Upon receipt, we extract the audio component of the media file. While any of the models summarised in this work could be used to analyse the audio file, we opt for deploying one of the baseline approaches: the extraction of ComParE features combined with an SVM classifier. For the current deployment, the audio file is first segmented into acoustic frames of 1 sec length and 25% overlap. Each frame is analysed with the model individually, and the final result is obtained through a majority voting scheme across frames to identify if the speaker is wearing a mask. This information is then transferred back to the smartphone app to be displayed to the user. If the model infers that the current user is wearing a face mask, the message of Fig. 7b is displayed on the screen; otherwise, the one of Fig. 7c.

Furthermore, we benchmark CNN architectures ResNet101 [21], DenseNet121 [20], and VGG19 [19] (used by the top participants [54]) by deploying each model using the Docker setup of TensorFlow serving[10] on a device with an Intel(R) Core(TM) i7-8700PU @ 3.20GHz CPU, an Nvidia RTX 2080 GPU, and 64 GB RAM memory. In order to make use of the parallelisation due to batching, we configure the server to wait for 1 ms, and group all received requests into batches (of at most 64 frames) for inference. In Table 3, we measure latency and throughput on both GPU and CPU-only, also using one 1 s frame per request or a track of 16 frames per request (corresponding to a 12 s track with the 25% overlap specified earlier). We do not include preprocessing time of calculating spectrograms. Similar to [54], we use an image of dimensions $320 \times 320 \times 3$ corresponding to the multi-channel spectrogram of a 1 s frame. Requests are sent on the same local host of the server, in order to diminish the arbitrary effects of network delays. Latency is measured by sending 100 requests sequentially, and measuring the average time until a response is returned. Throughput is measured by sending 100 requests concurrently and the total time used is measured after all of them are processed; we repeat this 10 times and report the average. Deploying the best ensemble of models used by Szep and Hariri [54] would slow down these values by a factor of 60 (20 instances of each of the 3 models), and would need much more GPU memory. The dedicated server can be avoided altogether by deploying an offline model directly on the app. However, we did not implement this, because it is computationally costly for the users without a dedicated mobile GPU, as seen by comparing CPU against GPU.

6. Conclusion

In this paper, we summarised the findings of the Mask Sub-Challenge (MSC) from the INTERSPEECH ComParE challenge series. The goal of this challenge was to develop techniques for predicting whether a speaker is wearing a surgical mask or not, based on audio only. This task gained momentum in the current COVID-19 pandemic, because of the sudden emergence of challenges to existing biometric techniques such as fingerprint recognition and facial detection, due to the virus's survival on surfaces and the mask mandates worldwide. These challenges placed a renewed focus on voice biometrics, which are less invasive.

First, we introduced the MASC database (used by MSC), which consists of audio recordings of people speaking both structured and unstructured text while wearing surgical masks or not wearing them. To review the submitted approaches, we introduced a general analysis framework, wherein we classified the methodologies of the presented approaches. The approaches mostly followed one of two patterns: Making use of spectrogram representations of audio in combination with several pretrained CNNs, or using audio features, which include the MSC baseline features as well as phonetic features. Most approaches made use of ensembles of several different models, which led to improved performance due to the diversity of the methods used. Multiple approaches attempted to artificially increase the size of the training data via data augmentation, training with GANs, or using Mixup. Furthermore, we presented an analysis of these approaches along two axes, *specific* against *generic* characteristics as well as whether they were based more on *algorithms* such as deep neural networks or more on hand-crafted (mostly phonetic) *features*. We then presented a fusion of the top approaches: Fusing the top five participants led to the best results. The presented advances could be of benefit to future voice biometrics approaches.

We discussed several aspects of the presented approaches. In particular, three key ingredients are necessary for the success of the models, namely ensemble learning, transfer learning, and data generation (mostly by using data augmentation) – all top models incorporated at least two of those in some form. The results obtained show that there is indeed a difference between using a mask and not using a mask for speech processing, which suggests an impact on voice biometrics.

This impact will likely be higher in realistic scenarios – note that the MASC scenario was optimal for recordings (clean controlled environment, high-quality microphones and no background noise). We discussed a few practical aspects of the models and some limitations, as well as the suitability of the presented models to voice biometrics. Furthermore, we presented a proof of concept Android app for smartphones. Together with the aforementioned results, this app motivates voice biometrics applications that can benefit from the classification task at hand; it can improve the application accordingly by automatically choosing a suitable speaker identification model based on the mask-wearing prediction. Also, we benchmarked the run-time of the top participants' models, if they are to be used for serving.

Finally, the findings suggest that applications can start to rely more on voice biometrics in the future, especially with the regulations of wearing masks.

[10] https://www.tensorflow.org/tfx/serving/docker, visited on 15.Aug.2021

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

References

[1] J. Wayman, A Definition of Biometrics National Biometric Test Center Collected Works 1997–2000, San Jose State University (2000).
[2] J. Wayman, A. Jain, D. Maltoni, D. Maio, An introduction to Biometric Authentication systems, Springer London, 2005.
[3] N. van Doremalen, T. Bushmaker, D.H. Morris, M.G. Holbrook, A. Gamble, B.N. Williamson, A. Tamin, J.L. Harcourt, N.J. Thornburg, S.I. Gerber, J.O. Lloyd–Smith, E. de Wit, V.J. Munster, Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1, N top N. Engl. J. Med. (2020) 1564–1567.
[4] D.K. Chu, E.A. Akl, S. Duda, K. Solo, S. Yaacoub, H.J. Schünemann, D.K. Chu, et al., Physical distancing, face masks, and eye protection to prevent person–to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis, The Lancet (2020) 1973–1987.
[5] M. Gomez-Barrero, P. Drozdowski, C. Rathgeb, J. Patino, M. Todisco, A. Nautsch, N. Damer, J. Priesnitz, N. Evans, C. Busch, Biometrics in the Era of COVID-19: Challenges and Opportunities, 2021,
[6] R. Saeidi, T. Niemi, H. Karppelin, J. Pohjalainen, T. Kinnunen, P. Alku, Speaker Recognition For Speech Under Face Cover, in: Proceedings INTERSPEECH, ISCA, Dresden, Germany, 2015, pp. 1012–1016.
[7] F. Sigona, Voice biometrics technologies and applications for healthcare: an overview, JDReAM. Journal of interDisciplinary REsearch Applied to Medicine (2018) 5–16.
[8] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Commun (1995) 91–108.
[9] M. McLaren, L. Ferrer, D. Castan, A. Lawson, The Speakers in the Wild (SITW) Speaker Recognition Database, in: Proceedings INTERSPEECH, ISCA, San Francisco, CA, USA, 2016, pp. 818–822.
[10] R.H. Woo, A. Park, T.J. Hazen, The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments, in: Odyssey - The Speaker and Language Recognition Workshop, IEEE, San Juan, PR, USA, 2006, pp. 1–6.
[11] A. Boles, P. Rad, Voice Biometrics: Deep Learning-based Voiceprint Authentication System, in: 12th System of Systems Engineering Conference (SoSE), IEEE, Waikoloa, HI, USA, 2017, pp. 1–6.
[12] K.J. Wittum, L. Feth, E. Hoglund, The effects of surgical masks on speech perception in noise, J. Acoust. Soc. Am. (2013). 060125–060125
[13] L.L. Mendel, J.A. Gardino, S.R. Atcherson, Speech understanding using surgical masks: A Problem in health care? J Am Acad Audiol (2008) 686–695.
[14] C. Llamas, P. Harrison, D. Donnelly, D. Watt, Effects of different types of face coverings on speech acoustics and intelligibility, York Papers in Linguistics Series 2 (2008) 80–104.
[15] T. Kawase, K. Yamaguchi, T. Ogawa, K. ichi Suzuki, M. Suzuki, M. Itoh, T. Kobayashi, T. Fujii, Recruitment of fusiform face area associated with listening to degraded speech sounds in auditory-visual speech perception: a PET study, Neurosci. Lett. (2005) 254–258.
[16] D.D. Nguyen, P. McCabe, D. Thomas, A. Purcell, M. Doble, D. Novakovic, A. Chacon, C. Madill, Acoustic voice characteristics with and without wearing a facemask, Sci Rep (2021) 1–11.
[17] R.M. Corey, U. Jones, A.C. Singer, Comparison of the acoustic effects of face masks on speech, Hear J (2021) 36–38.
[18] Ö.D. Orman, L.M. Arslan, Frequency Analysis of Speaker Identification, in: A Speaker Odyssey-The Speaker Recognition Workshop, ISCA, Crete, Greece, 2001, pp. 219–222.
[19] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014.
[20] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 4700–4708.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.
[22] S. Hershey, S. Chaudhuri, D.P. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, M. Seybold Bryan Slaney, R.J. Weiss, K. Wilson, CNN Architectures for Large-Scale Audio Classification, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, New Orleans, LA, USA, 2017, pp. 131–135.
[23] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, et al., Deep Speech 2 : End-to-end speech recognition in English and Mandarin, in: Proceedings of The 33rd International Conference on Machine Learning, PMLR, New York, NY, USA, 2016, pp. 173–182.
[24] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, K.R. Scherer, On the acoustics of emotion in audio: what speech, music and sound have in common, Frontiers in Emotion Science (2013) 1–12.
[25] F. Eyben, Real-time speech and music classification by large audio feature space extraction, Springer International Publishing, 2015.
[26] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, Q. Sun, Deep learning for image-based cancer detection and diagnosis – A survey, Pattern Recognit 83 (2018) 134–149.
[27] J. Shuja, E. Alanazi, W. Alasmary, A. Alashaikh, COVID-19 Open source data sets: a comprehensive survey, Applied Intelligence (2020) 1–30.
[28] G. Deshpande, B. Schuller, An Overview on Audio, Signal, Speech, & Language Processing for COVID-19, 2020.
[29] T. Alafif, A. Tehame, S. Bajaba, A. Barnawi, S. Zia, Machine and deep learning towards COVID-19 diagnosis and treatment: survey, challenges, and future directions, Int J Environ Res Public Health (2021) 1117.
[30] M.A. Nessiem, M.M. Mohamed, H. Coppock, A. Gaskell, B.W. Schuller, Detecting COVID-19 from breathing and coughing sounds using deep neural networks, in: 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, Lisbon, Portugal, 2021, pp. 183–188.
[31] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, C. Mascolo, Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3474–3484.
[32] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, Biometrics recognition using deep learning: A survey, 2019.
[33] K. Sundararajan, D.L. Woodard, Deep Learning for Biometrics: A Survey, ACM Computing Surveys (CSUR) (2018) 1–34.
[34] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognit (2018) 354–377.
[35] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, B. Schuller, Snore Sound Classification Using Image-based Deep Spectrum Features, in: Proceedings INTERSPEECH, ISCA, Stockholm, Sweden, 2017, pp. 3512–3516.
[36] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M.D. Plumbley, PANNs: Large-Scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Trans Audio Speech Lang Process (2020) 2880–2894.
[37] B.W. Schuller, A.M. Batliner, Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing, Wiley, 2014.
[38] B.W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A.D. MacIntyre, S. Hantke, The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2042–2046.
[39] C.D. Manning, P. Raghavan, H. Schütze, An introduction to information retrieval, Cambridge University Press, 2009.
[40] A.K. Jain, A.A. Ross, K. Nandakumar, Introduction to biometrics, Springer US, Boston, MA, USA, 2011.
[41] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2009 Emotion Challenge, in: Proceedings of INTERSPEECH, ISCA, Brighton, England, 2009, pp. 312–315.
[42] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim, The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH, ISCA, Lyon, France, 2013, pp. 148–152.
[43] F. Eyben, F. Weninger, F. Groß, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, in: Proceedings of the ACM International Conference on Multimedia, Association for Computing Machinery, Barcelona, Spain, 2013, pp. 835–838.
[44] M. Schmitt, B. Schuller, OpenXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit, Journal of Machine Learning Research (2017) 1–5.
[45] H. Lim, M.J. Kim, H. Kim, Robust sound event classification using LBP-HOG based Bag-of-Audio-Words feature representation, in: Proceedings INTERSPEECH, ISCA, Dresden, Germany, 2015, pp. 3325–3329.
[46] S. Amiriparian, Deep Representation Learning Techniques for Audio Signal Processing, Technische Universität München, 2019 Ph.D. thesis.
[47] S. Amiriparian, M. Schmitt, S. Ottl, M. Gerczuk, B. Schuller, Deep Unsupervised Representation Learning for Audio-based Medical Applications, Springer, 2020.
[48] S. Amiriparian, M. Freitag, N. Cummins, B. Schuller, Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio, in: Proceedings DCASE, Tampere University of Technology. Laboratory of Signal Processing, Munich, Germany, 2017, pp. 17–21.
[49] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, B. Schuller, Audeep: unsupervised learning of representations from audio with deep recurrent neural networks, Journal of Machine Learning Research (2018) 1–5.
[50] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition, in: Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, Geneva, Switzerland, 2013, pp. 511–516.
[51] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: a simple data augmentation method for automatic speech recognition, in: Proceedings INTERSPEECH, ISCA, Graz, Austria, 2019, pp. 4110–4114.
[52] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization, 2018.

[53] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[54] J. Szep, S. Hariri, Paralinguistic classification of mask wearing by image classifiers and fusion, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2087–2091.
[55] C. Montacié, M.-J. Caraty, Phonetic, frame clustering and intelligibility analyses for the INTERSPEECH 2020 ComParE challenge, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2062–2066.
[56] T. Koike, K. Qian, B.W. Schuller, Y. Yamamoto, Learning higher representations from pre-trained deep models with data augmentation for the ComParE 2020 challenge mask task, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2047–2051.
[57] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, A. Karpov, Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 Mask and Breathing Sub-Challenges, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2072–2076.
[58] P. Klumpp, T. Arias-Vergara, J.C. Vásquez-Correa, P.A. P'erez-Toro, F. Hönig, E. Nöth, J.R. Orozco-Arroyave, Surgical mask detection with deep recurrent phonetic models, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2057–2061.
[59] Z. Yang, Z. An, Z. Fan, C. Jing, H. Cao, Exploration of Acoustic and Lexical Cues for the INTERSPEECH 2020 Computational Paralinguistic Challenge, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2092–2096.
[60] N.-C. Ristea, R.T. Ionescu, Are you wearing a mask? Improving mask detection from speech using augmentation by cycle-consistent GANs, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2102–2106.
[61] S. Illium, R. Müller, A. Sedlmeier, C. Linnhoff-Popien, Surgical mask detection with convolutional neural networks and data augmentations on spectrograms, in: Proceedings INTERSPEECH, ISCA, Shanghai, China, 2020, pp. 2052–2056.
[62] H. Wu, L. Zhang, L. Yang, X. Wang, J. Wang, D. Zhang, M. Li, Mask Detection and Breath Monitoring from Speech: on Data Augmentation, Feature Representation and Modeling, 2020.
[63] X. Xu, J. Deng, Z. Zhang, C. Wu, B. Schuller, Identifying surgical-mask speech using deep neural networks on low-level aggregation, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, Association for Computing Machinery, Virtual Event, Republic of Korea, 2021, pp. 580–585.
[64] A. Chan, E. Gouvea, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines, M. Seltzer, The Hieroglyphs: building speech applications using CMU Sphinx and related resources, 2007.
[65] L. Breiman, Random forests, Mach Learn (2001) 5–32.
[66] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory and practice, Int J Comput Vis (2013) 222–245.
[67] E. Isaac, Test of Hypothesis - Concise Formula Summary, 2015, Ms.
[68] R.L. Wasserstein, N.A. Lazar, The ASA's statement on *p*-values: context, process, and purpose, Am Stat (2016) 129–133.
[69] A. Batliner, S. Hantke, B. Schuller, Ethics and good practice in computational paralinguistics, Transactions on Affective Computing (2020). 1–1
[70] S. Nzuva, L. Nderu, The superiority of the ensemble classification methods: A Comprehensive review, Journal of Information Engineering & Applications (2019) 43–53.
[71] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge, Speech Commun (2011) 1062–1087.
[72] F. Cen, X. Zhao, W. Li, G. Wang, Deep feature augmentation for occluded image classification, Pattern Recognit (2021) 107737.
[73] G. Li, M. Zhang, J. Li, F. Lv, G. Tong, Efficient densely connected convolutional neural networks, Pattern Recognit (2021) 107610.
[74] P. Mohan, A.J. Paul, A. Chirania, A tiny CNN architecture for medical face mask detection for resource-constrained endpoints, in: Innovations in Electrical and Electronic Engineering, Springer Singapore, 2021, pp. 657–670.

Mostafa M. Mohamed received his M.Sc. degree in Computer Science at the University of Freiburg, Germany. Currently, he is an external Ph.D. student at the University of Augsburg, and a Senior Research Data Scientist at SyncPilot GmbH. His main research interests are in applying deep learning to various applications, including Speech Enhancement applications.

Mina A. Nessiem received his M.Sc. degree in Computer Science at the University of Augsburg. He is an external Ph.D. student at the University of Augsburg and a Senior Research Data Scientist at SyncPilot GmbH. His interests generally relate to the usage of deep learning within industrial applications and specifically within the field of affective computing.

Anton Batliner is with EIHW, University of Augsburg. He is co-editor/author of two books and author/co-author of more than 300 technical articles, with an *h*-index = 48 and > 11 000 citations. His main research interests are all (cross-linguistic) aspects of prosody and (computational) paralinguistics.

Christian Bergler is a Ph.D. student at Friedrich-Alexander-University (FAU) Erlangen-Nuremberg. His research is focused on machine learning applied to the field of bioacoustics, in order to analyse animal recordings, in particular killer whale underwater signals, to identify significant communication patterns, correlating them to the respective animal behaviour in order to decode animal communication.

Simone Hantke received her Ph.D. degree in 2019 from the Technische Universität München (TUM). Her interests lie in the field of affective computing and speech recognition, focusing on data collection and new machine learning approaches. She is currently working as the Lead Project Manager for Data Intelligence at audEERING.

Maximilian Schmitt is a Ph.D. student at EIHW, University of Augsburg. He received his diploma degree in Electrical Engineering from RWTH Aachen University. His research is focused on signal processing, intelligent audio analysis, machine learning, and crossmodal affect recognition.

Alice Bairdreceived her M.F.A. degree from Columbia University's Computer Music Centre, and is currently a Ph.D. student at EIHW, University of Augsburg. Her research is focused on intelligent audio analysis in the domain of both speech and general audio.

Adria Mallol-Ragolta received his M.Sc. degree in Electrical Engineering at the University of Colorado, Colorado Springs, USA, and is currently a Ph.D. student at EIHW, University of Augsburg. His research interests include the computational understanding of human affect and health states using multimodal and ubiquitous computing solutions.

Vincent Karas received his M.Sc. degree in Medical Engineering at Technische Universität München (TUM). He is currently a Ph.D. student at EIHW, University of Augsburg, and a member of the Ph.D. candidate program at BMW. His research focuses on multimodal affect recognition.

Shahin Amiriparian has received his Ph.D. degree with highest honours (summa cum laude) from Technische Universität München (TUM). Currently, he is a habilitation candidate at EIHW, University of Augsburg. His main research focus is deep learning, unsupervised representation learning, and transfer learning for machine perception, affective computing, and audio understanding.

Björn W. Schuller received his diploma, Ph.D. degree, habilitation, and Adjunct Teaching Professor in Machine Intelligence and Signal Processing, all in EE/IT from TUM in Munich/Germany. He is Full Professor of Artificial Intelligence and the Head of GLAM at Imperial College London/UK, Full Professor and Chair of EIHW, University of Augsburg, co-founding CEO and current CSO of audEERING.