IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Face Mask Type and Coverage Area Recognition from Speech with Prototypical Networks

Adria Mallol-Ragolta[1,3], *Anika A. Spiesberger*[1,3], *Björn W. Schuller*[1,2,3,4]

[1] CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany
[2] MDSI – Munich Data Science Institute, Germany
[3] MCML – Munich Center for Machine Learning, Germany
[4] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

adria.mallol-ragolta@tum.de

Abstract

We investigate the use of prototypical networks on the problems of face mask type (3 classes), face mask coverage area (3 classes), and face mask type and coverage area (5 classes) recognition from speech. We explore the MASCFLICHT Corpus, a dataset containing 2 h 27 m 55 s of speech data from 30 German speakers recorded with a smartphone. We extract formant-related features and the spectrogram representations from the samples. We enrich the spectrograms overlaying the traces of the central frequency of the first four formants. Our experiments also consider the fusion via concatenation of the embedded representations extracted from the formant-related features and the spectrogram representations. We implement classification- and prototypical encoder-based networks. The results obtained on the test sets support the suitability of the prototypical encoder models, scoring an Unweighted Average Recall (UAR) of 49.9 %, 45.0 %, and 31.6 % on the three considered problems, respectively.
Index Terms: Face Mask Recognition, Paralinguistic Analysis, Prototypical Networks, Digital Health.

1. Introduction

Face masks are an effective instrument to prevent the transmission of *Respiratory Viral Infections* (RVI) [1]. During the *Coronavirus Disease 2019* (COVID-19) pandemic, for instance, face masks became a key instrument to control the spread of the disease [2] and reduce the number of COVID-19 deaths [3]. To protect patients and the people in their surroundings – with a special emphasis on the healthcare professionals – we envision the development of face mask monitoring tools that can help control the droplet transmission of respiratory diseases [4, 5].

The computer vision community has vastly investigated the face mask detection problem [6, 7, 8, 9, 10, 11, 12, 13]. In the computer audition domain, the research on this problem was motivated by the release of the *Mask Augsburg Speech Corpus* (MASC) as part of the INTERSPEECH 2020 *Computational Paralinguistics ChallengE* (ComParE) [14, 15]. Exploiting this dataset, researchers analysed the speech changes produced by wearing a mask [16, 17], used deep neural networks on the spectrogram or the Mel-spectrogram representations of the audio signals [18, 19], and investigated transfer learning-based solutions [20, 21]. As the MASC dataset only contains speech samples from participants wearing a surgical face mask, researchers revisited the problem, expanding the face mask types worn by the participants [22], and considering not only different face mask types, but also different coverage areas [23]; i. e., with the mask covering only the mouth, or both the mouth and the nose.

Inspired by related works utilising prototypical networks for aphasia [24] and snore sounds classification [25], we employ prototypical networks in the face mask type and coverage area recognition from speech with the hypothesis that prototypical representations can help outperform traditional classification approaches. We exploit the *Mask Augsburg Speech Corpus using FFP2 and surgicaL masks with the aIrways Covered Halfway and compleTely* (MASCFLICHT) [23]. We extract formant-related features and the spectrogram representations from the speech samples. We also aim to investigate the suitability of enriching the spectrograms with the traces corresponding to the central frequency of the first four formants, so that the spectrograms not only capture the frequencies and their corresponding energy levels in the voice, but also carry vocal tract information. We employ the *Unweighted Average Recall* (UAR) as the evaluation metric to assess model performances.

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

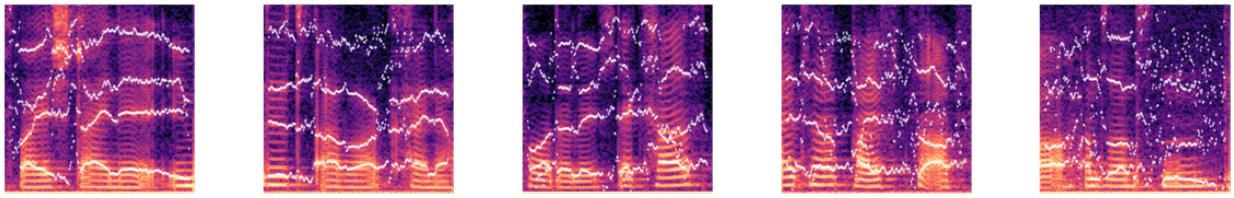
2. The MASCFLICHT Corpus

The MASCFLICHT Corpus [23] contains 2 h 27 m 55 s of speech samples from 30 German participants (15 f, 15 m) recorded with the microphone embedded in a Xiaomi Mi 10 smartphone. Participants' voice was recorded under five different conditions: i) without a 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). The surgical face mask from LyncMed, and the FFP2 face mask from IPOS were employed to collect the data. Under each condition, study participants described a picture (free speech task), and read ten sentences (guided speech task). Participants' habituation to the proposed tasks was overcome by creating five different, yet equivalent exercises for each task. The exercise-condition pairs were randomised for each participant, so that all the exercises were recorded under all the five aforementioned conditions. The statistics of the MASCFLICHT Corpus are summarised in Table 1. The dataset includes 8 875 speech samples of 1 s length distributed in three speaker-independent partitions – train (5 606), devel (1 568), and test (1 701). The samples are stored in mono, 16-bit PCM format, and with a sampling rate of 16 kHz.

3. Methodology

This section introduces the methodology followed. Section 3.1 describes the representations explored from the speech samples, Section 3.2 details the neural networks implemented, and Section 3.3 defines the parameters utilised for training the models.

10.21437/IberSPEECH.2024-27



(a) *No Mask.* (b) *Surgical Partial.* (c) *Surgical Total.* (d) *FFP2 Partial.* (e) *FFP2 Total.*

Figure 1: *Spectrogram visualisations from speech samples belonging to the same speaker recorded under the five different conditions investigated with the estimated four first formants overlaid.*

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

3.1. Data Representations

We investigate the use of formant-related features and spectrograms as representations to extract from the available speech samples. We employ the `parselmouth` library [26], v.0.4.3, to extract the central frequency and the corresponding bandwidth ($F_C \oplus F_B$) of the first four formants with a hop size of 64 samples (4 ms). With these parameters, an 8-dimensional sequence of formant-related features characterises each speech sample. The spectrogram representations (S) are computed utilising the `librosa` library [27], v.0.8.0. Specifically, we first compute the magnitude of the *Short-Time Fourier Transform* (STFT) of each speech sample using a window length of 512 samples (32 ms), and a hop size of 64 samples (4 ms). We then store the generated spectrograms as colour images of 224×224 pixels, previously converted to dB, and min-max normalised. The defined hop sizes ensure that both modalities have the same resolution in the time domain. This is relevant for the third representation of the speech samples we exploit, which overlays the central frequency of the first four formants in the spectrogram representations (S w/ F_C). With this approach, we aim to enrich the spectrograms, as they not only contain information related to the frequencies in the human voice and their corresponding energy levels, but also include characteristics of the vocal tract. The enriched spectrograms under the five investigated conditions (cf. Section 2) are showcased in Figure 1.

3.2. Models Description

We tackle the proposed tasks in the MASCFLICHT Corpus as classification problems. The models we implement are composed of two blocks: an encoder block, and a classification block. While the former is responsible for learning deep-learnt representations from the input modalities, the latter performs the actual classification. The encoder block generates a 32-dimensional representation at the output, and its implementation depends on the actual features to model. To favour the comparison, the implementation of the classification block in the unimodal models remains unaltered, featuring a dropout layer with a probability of 60 % and a single linear layer with 32 input neurons and as many neurons at the output as classes we aim our model to recognise.

The encoder block when modelling the formant-related features unimodally ($\text{Clf}^{\text{F}}_{\text{CNN}}$) implements a 1-dimensional convolutional layer with 8 and 128 input and output channels, respectively, and a kernel size of 5. Following the convolutional layer, we use batch normalisation, and the output is transformed with a *Rectified Linear Unit* (ReLU) function. We then implement

Table 2: *Summary of the number of trainable parameters for each model investigated in the face mask type (FM Type), face mask coverage area (FM Coverage), and face mask type and coverage area (FM Type & Coverage) recognition problems, according to the different input modalities explored.*

Features	**Model Arch.**	**FM Type**	**FM Coverage**	**FM Type & Coverage**
$F_C \oplus F_B$	$\text{Clf}^{\text{F}}_{\text{CNN}}$	22 019	22 019	22 085
	Proto. Naïve	–	–	–
	Proto. $\text{Enc}^{\text{F}}_{\text{CNN}}$	21 920	21 920	21 920
S	$\text{Clf}^{\text{S}}_{\text{PT}}$	16 515	16 515	16 581
	$\text{Clf}^{\text{S}}_{\text{FT}}$	11 193 027	11 193 027	11 193 093
or	Proto. Naïve	–	–	–
S w/ F_C	Proto. $\text{Enc}^{\text{S}}_{\text{PT}}$	16 416	16 416	16 416
	Proto. $\text{Enc}^{\text{S}}_{\text{FT}}$	11 192 928	11 192 928	11 192 928
$S \oplus F_C \oplus F_B$	$\text{Clf}^{\text{F, S}}_{\text{CNN, PT}}$	38 531	38 531	38 661
	$\text{Clf}^{\text{F, S}}_{\text{CNN, FT}}$	11 215 043	11 215 043	11 215 173
	Proto. Naïve	–	–	–
	Proto. $\text{Enc}^{\text{F, S}}_{\text{CNN, PT}}$	38 336	38 336	38 336
	Proto. $\text{Enc}^{\text{F, S}}_{\text{CNN, FT}}$	11 214 848	11 214 848	11 214 848

Table 3: *Performance summary of the face mask type (FM Type), face mask coverage area (FM Coverage), and face mask type and coverage area (FM Type & Coverage) recognition models trained in terms of the UAR (%). We include the CI on the test partition, and the optimal model training time.*

Features	Model Arch.	FM Type				FM Coverage				FM Type & Coverage			
		UAR [%]			Tr. Time	UAR [%]			Tr. Time	UAR [%]			Tr. Time
		Dev	Test	CI on Test	(MM:)SS	Dev	Test	CI on Test	(MM:)SS	Dev	Test	CI on Test	(MM:)SS
$F_C \oplus F_B$	Clf^{F}_{CNN}	40.8	40.0	(38.0 – 42.0)	6:02	42.6	38.1	(36.0 – 40.4)	2:19	28.8	27.9	(25.9 – 30.0)	5:56
	Proto. Naïve	43.0	38.9	(36.6 – 41.4)	13	41.0	38.6	(36.2 – 41.1)	13	26.8	24.6	(22.7 – 26.5)	13
	Proto. Enc^{F}_{CNN}	42.5	40.4	(38.2 – 42.7)	53	43.3	36.7	(34.5 – 39.1)	55	28.4	24.9	(23.1 – 26.8)	2:22
S	Clf^{S}_{PT}	39.8	40.0	(38.3 – 41.6)	8:51	41.5	38.5	(36.5 – 40.3)	18:27	31.8	**31.9**	(29.8 – 33.9)	22:03
	Clf^{S}_{FT}	55.1	46.7	(44.3 – 49.3)	54:23	52.0	42.9	(40.6 – 45.3)	2:15	34.1	30.5	(28.5 – 32.4)	4:32
	Proto. Naïve	43.2	40.8	(38.2 – 43.1)	3	42.4	41.9	(39.4 – 44.5)	3	25.4	28.0	(25.8 – 30.1)	3
	Proto. Enc^{S}_{PT}	44.1	41.7	(39.5 – 44.1)	2:26	44.4	38.6	(36.2 – 41.1)	2:01	31.0	31.3	(29.3 – 33.4)	19:51
	Proto. Enc^{S}_{FT}	55.1	**49.9**	(47.4 – 52.4)	16:27	51.4	38.7	(36.7 – 40.6)	10:07	36.0	**31.6**	(29.5 – 33.8)	21:58
S w/ F_C	Clf^{S}_{PT}	40.3	39.4	(37.9 – 41.2)	11:55	38.3	36.9	(35.0 – 38.7)	14:27	30.3	26.9	(25.0 – 28.9)	21:59
	Clf^{S}_{FT}	49.8	42.5	(40.6 – 44.4)	14:15	50.1	40.5	(38.1 – 42.8)	6:17	32.3	24.4	(23.0 – 25.7)	8:07
	Proto. Naïve	39.2	39.1	(36.8 – 41.6)	3	40.3	40.9	(38.4 – 43.5)	3	24.5	25.1	(23.2 – 27.2)	3
	Proto. Enc^{S}_{PT}	44.1	40.9	(38.6 – 43.5)	3:17	41.6	37.0	(34.6 – 39.4)	2:50	29.9	27.9	(25.8 - 29.9)	15:43
	Proto. Enc^{S}_{FT}	50.1	47.6	(45.2 – 49.9)	17:58	46.0	32.4	(30.6 – 34.2)	9:15	32.0	27.2	(25.6 – 28.7)	11:20
$S \oplus F_C \oplus F_B$	$Clf^{F,S}_{CNN,PT}$	43.1	37.1	(35.5 – 38.7)	2:21	47.3	39.3	(37.4 – 41.2)	41:55	30.2	29.7	(27.8 – 31.7)	10:23
	$Clf^{F,S}_{CNN,FT}$	54.9	46.6	(44.2 – 48.9)	20:06	48.6	42.6	(40.4 – 45.0)	32:25	35.2	25.1	(24.1 – 26.2)	3:35
	Proto. Naïve	43.0	38.9	(36.6 – 41.4)	16	41.0	38.6	(36.2 – 41.1)	16	26.8	24.6	(22.7 – 26.5)	16
	Proto. $Enc^{F,S}_{CNN,PT}$	46.3	43.2	(41.0 – 45.5)	23:04	44.8	43.4	(41.0 – 45.8)	12:59	30.6	31.4	(29.3 – 33.6)	67:35
	Proto. $Enc^{F,S}_{CNN,FT}$	52.9	48.5	(45.8 – 51.0)	56:38	50.4	**45.0**	(42.5 – 47.3)	52:56	34.4	29.1	(27.0 – 31.2)	42:44
Baseline [23]		49.2	49.3	–	–	53.8	47.8	–	–	31.9	35.0	–	–
Chance Level		33.3	33.3	–	–	33.3	33.3	–	–	20.0	20.0	–	–

a 1-dimensional adaptive average pooling layer, so that each channel produces four features at the output. Finally, we reduce the dimensionality of the embedded representations from 512 to 32 with a linear layer and a dropout of 60 %. We utilise the ResNet18 network [28] without the last layer to extract deep-learnt features from the spectrogram representations of the speech samples with or without the central frequency of the first four formants overlaid. The ResNet18 network produces a 512-dimensional representation of the input spectrograms, and we also reduce its dimensionality from 512 to 32 with a linear layer and a dropout of 60 %. We analyse the performance of this encoder block in two scenarios: utilising the pre-trained weights of ResNet18 (Clf^{S}_{PT}), and fine-tuning them (Clf^{S}_{FT}) for our tasks. To complement our analysis, we consider combining the formant-related features and the spectrogram representations multimodally by concatenating the embedded features the corresponding encoder blocks extract. This multimodal embedded representation is then fed into the classification block of the network, which, in this case, implements 64 input neurons. The spectrogram-dedicated encoder block in this multimodal scenario is also explored when fine-tuning ($Clf^{F,S}_{CNN,FT}$) the ResNet18 weights, and when using them pre-trained ($Clf^{F,S}_{CNN,PT}$).

The main goal of this work is to investigate the performance of prototypical networks [29] on the proposed tasks of the explored dataset. These networks aim at learning an embedded, prototypical representation of each targeted class. At inference time, the closest prototype to the embeddings extracted from the unseen samples determines the class assigned to the latter. The prototypical networks operate directly at the 32-dimensional output of the unimodal encoder blocks and at the 64-dimensional output of the multimodal encoder blocks. We compute the prototypical representation of each targeted class as:

$$\mathbf{c}_k = \frac{1}{\mid S_k \mid} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi\left(\mathbf{x}_i\right), \tag{1}$$

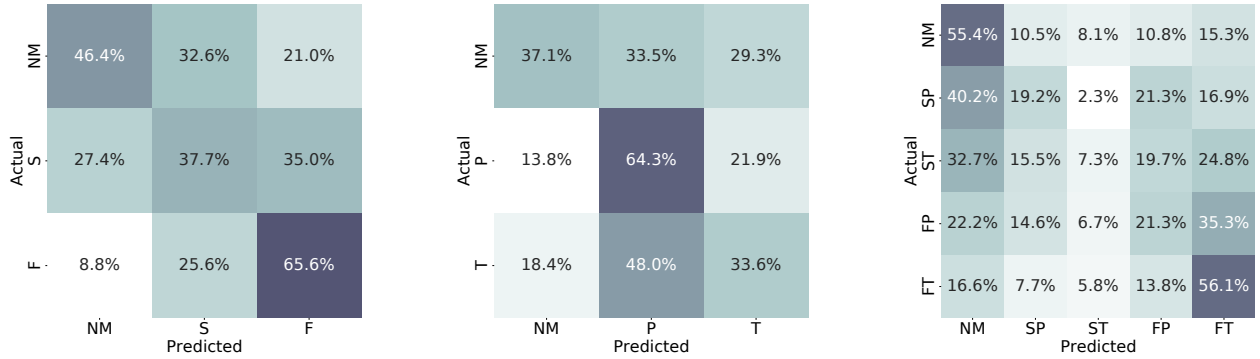
where $f_\phi\left(\cdot\right)$ represents the embedded representation learnt at the output of the encoder block. We normalise the prototypes so they are unit-length, as we have experimentally observed that this benefits the model performances. We train the prototypical encoders (Proto. Enc) by minimising the prototypical loss:

$$\mathcal{L}\left(\theta\right) = -\log\left(\frac{\exp\left(-d\left(f_\phi\left(\mathbf{x}\right), \mathbf{c}_k\right)\right)}{\sum_{k'} \exp\left(-d\left(f_\phi\left(\mathbf{x}\right), \mathbf{c}_{k'}\right)\right)}\right). \tag{2}$$

This loss aims at minimising the distance between the embeddings of the current sample and its corresponding prototype, and maximising the distance with the other prototypes. We employ the Euclidean distance as the distance function $d\left(\cdot\right)$ in Equation (2). We investigate this approach with the different encoder blocks defined previously. As a baseline, we include a naïve prototypical model (Proto. Naïve), which averages the corresponding representations from the targeted classes without training. In the case of the formant-related features, we average the sequence of features extracted per sample before computing the prototypes. When working with the spectrograms or the enriched spectrograms, we compute the prototypes utilising the features extracted with the pre-trained ResNet18 model.

3.3. Models Training

All the models are trained under the exact same conditions for a fair comparison. Table 2 summarises the number of trainable



(a) *NM: No Mask, S: Surgical mask, F: FFP2 mask.*

(b) *NM, P: Partial coverage (only mouth), T: Total coverage (mouth and nose).*

(c) *NM, SP: Surgical Partial, ST: Surgical Total, FP: FFP2 Partial, FT: FFP2 Total.*

Figure 2: *Confusion matrices of the best performing face mask type (Figure 2a), face mask coverage area (Figure 2b), and face mask type and coverage area (Figure 2c) recognition models assessed on the test set.*

parameters each model requires. The classification-based models are trained by minimising the categorical cross-entropy loss, while the prototypical-based models minimise the prototypical loss defined in Equation (2). Regardless of the loss function, we use Adam as the optimiser with a fixed learning rate of 10^{-3}. We select the 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 512 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 to determine the optimal number of training epochs that minimises the risk of overfitting. The models are implemented with the `PyTorch` library, v.1.7.0, seeding the pseudorandom number generator at the models' initialisation.

4. Experimental Results

This section presents the results obtained (cf. Table 3). The model performances are reported in terms of the UAR. We include the 95 % *Confidence Intervals* (CI) on the test set, computed by repeating 1 000 times the bootstrapping approach. We include the training time of the optimal model, which merges the samples belonging to the train and the development partitions.

Face Mask Type Recognition. The Proto. $\text{Enc}^{\text{S}}_{\text{FT}}$ model exploiting the spectrogram representations of the speech samples scores the best performance on the test set with a UAR of 49.9 %, surpassing the baseline performance [23], which obtained a UAR of 49.3 %. Analysing the confusion matrix of the Proto. $\text{Enc}^{\text{S}}_{\text{FT}}$ model (cf. Figure 2a), the main confusions take place between the samples recorded with no mask and the samples recorded with the surgical mask, and between the samples recorded with the surgical and the FFP2 masks. Analysing the numerical results (cf. Table 3) per modality, the prototypical encoder models obtain the highest UAR scores in all the cases investigated. This result supports the creation of prototypical representations from the targeted classes for the face mask type recognition problem. We also observe that the fine-tuned models exploiting the spectrogram representations outperform the pre-trained ones.

Face Mask Coverage Area Recognition. The multimodal Proto. $\text{Enc}^{\text{F,S}}_{\text{CNN,FT}}$ model fusing the formant-related features and the spectrogram representation achieves the best performance on the test set, with a UAR of 45.0 %. The resulting confusion matrix after comparing the ground truth and the inferences produced by this model is depicted in Figure 2b. From the inspection of the confusion matrix, we detect that 48.0 % of the samples recorded with the mouth and the nose covered are incorrectly classified as if they were recorded only with the mouth covered. This result indicates the difficulty of differentiating between these two classes from a computer audition perspective.

Face Mask Type and Coverage Area Recognition. The results reported in Table 3 indicate the appropriateness of the spectrogram modality in this problem. The $\text{Clf}^{\text{S}}_{\text{PT}}$ model scores the best UAR, 31.9 %, on the test set. The Proto. $\text{Enc}^{\text{S}}_{\text{FT}}$ model follows closely with a UAR of 31.6 %. Despite not achieving the best result, we can reasonably argue about the suitability of the Proto. $\text{Enc}^{\text{S}}_{\text{FT}}$ model. Figure 2c depicts the confusion matrix generated from the results obtained with the $\text{Clf}^{\text{S}}_{\text{PT}}$ model. The model is capable of recognising the samples recorded with no mask, and the samples recorded with an FFP2 mask with both the mouth and the nose covered. Nevertheless, the model underperforms in the recognition of the remaining categories.

5. Conclusions

We investigated the utilisation of prototypical networks on the face mask type, face mask coverage area, and face mask type and coverage area recognition problems exploiting the speech samples available in the MASCFLICHT Corpus. We also proposed enriching the spectrogram representations of the speech samples with the central frequency of the first four formants, so the spectrograms additionally contain information from the vocal tract. This visual fusion showed limited success in the results obtained. The prototypical encoders achieved promising results on the test sets of the three problems considered. In the face mask type recognition problem, a prototypical encoder exploiting the spectrogram representations scored a UAR of 49.9 %. In the face mask coverage area recognition problem, a prototypical encoder exploiting the multimodal fusion of formant-related features and spectrograms achieved a UAR of 45.0 %. In the face mask type and coverage area recognition problem, a prototypical encoder exploiting the spectrograms obtained the second highest UAR, 31.6 %, below a classification-based model, which scored a UAR of 31.9 %. In future works, we envision expanding the representations to extract from the speech samples to further assess the suitability of the proposed prototypical encoders in these problems. Follow-up studies could focus on the investigation of prototypical networks in encoder-decoder architectures to analyse the robustness of the generated prototypes.

6. Acknowledgements

This project has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), and from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101060660 (SHIFT).

7. References

[1] M. Liang, L. Gao, C. Cheng, Q. Zhou, J. P. Uy, K. Heiner, and C. Sun, "Efficacy of face mask in preventing respiratory virus transmission: A systematic review and meta-analysis," *Travel Medicine and Infectious Disease*, vol. 36, 2020, paper ID: 101751.

[2] 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.

[3] 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.

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

[5] G. Brankston, L. Gitterman, Z. Hirji, C. Lemieux, and M. Gardam, "Transmission of influenza A in human beings," *The Lancet. Infectious Diseases*, vol. 7, no. 4, pp. 257–265, 2007.

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

[7] 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 *Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies*. Kharagpur, India: IEEE, 2020, article ID 49239.

[8] 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 *Proceedings of the 6th International Conference on Computer and Communications*. Chengdu, China: IEEE, 2020, pp. 1398–1404.

[9] 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.

[10] 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.

[11] 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.

[12] C. Dewi, D. Manongga, Hendry, E. Mailoa, and K. D. Hartomo, "Deep Learning and YOLOv8 Utilized in an Accurate Face Mask Detection System," *Big Data and Cognitive Computing*, vol. 8, no. 1, 2024.

[13] S. Verma, P. Rani, S. Gupta, R. Sharma, K. Yadav, A. N. Aledaily, and M. Alharbi, "An automated face mask detection system using transfer learning based neural network to preventing viral infection," *Expert Systems*, vol. 41, 2024.

[14] 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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[15] 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.

[16] C. Montacié and M.-J. Caraty, "Phonetic, Frame Clustering and Intelligibility Analyses for the INTERSPEECH 2020 ComParE Challenge," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA, 2020, pp. 2062–2066.

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

[18] J. Szep and S. Hariri, "Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA, 2020, pp. 2087–2091.

[19] 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.

[20] 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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA, 2020, pp. 2047–2051.

[21] 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 *Proceedings of the 21st Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA, 2020, pp. 2072–2076.

[22] M. Markitantov, E. Ryumina, D. Ryumin, and A. Karpov, "Biometric Russian Audio-Visual Extended MASKS (BRAVE-MASKS) Corpus: Multimodal Mask Type Recognition Task," in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*. Incheon, Korea: ISCA, 2022, pp. 1756–1760.

[23] A. Mallol-Ragolta, N. Urbach, S. Liu, A. Batliner, and B. Schuller, "The MASCFLICHT Corpus: Face Mask Type and Coverage Area Recognition from Speech," in *Proceedings of the 24th Annual Conference of the International Speech Communication Association*. Dublin, Ireland: ISCA, 2023, pp. 2358–2362.

[24] M. Zusag, L. Wagner, and T. Bloder, "Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification," in *Proceedings of the 24th Annual Conference of the International Speech Communication Association*. Dublin, Ireland: ISCA, 2023, pp. 3013–3017.

[25] L. Ding and J. Peng, "Automatic classification of snoring sounds from excitation locations based on prototypical network," *Applied Acoustics*, vol. 195, 2022.

[26] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[27] 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 *Proceedings of the 14th Python in Science Conference*. Austin, TX, USA: SciPy, 2015, pp. 18–25.

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

[29] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*. Long Beach, CA, USA: NeurIPS, 2017, 11 pages.