2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871125

Triplet Loss-Based Models for COVID-19 Detection from Vocal Sounds

Adria Mallol-Ragolta[1*], Florian B. Pokorny[1,2], Katrin D. Bartl-Pokorny[1,2],
Anastasia Semertzidou[1], and Björn W. Schuller[1,3]

Abstract— **This work focuses on the automatic detection of COVID-19 from the analysis of vocal sounds, including sustained vowels, coughs, and speech while reading a short text. Specifically, we use the Mel-spectrogram representations of these acoustic signals to train neural network-based models for the task at hand. The extraction of deep learnt representations from the Mel-spectrograms is performed with Convolutional Neural Networks (CNNs). In an attempt to guide the training of the embedded representations towards more separable and robust inter-class representations, we explore the use of a triplet loss function. The experiments performed are conducted using the Your Voice Counts dataset, a new dataset containing German speakers collected using smartphones. The results obtained support the suitability of using triplet loss-based models to detect COVID-19 from vocal sounds. The best Unweighted Average Recall (UAR) of 66.5 % is obtained using a triplet loss-based model exploiting vocal sounds recorded while reading.**

I. INTRODUCTION

The pandemic caused by the outbreak of the *Coronavirus Disease 2019* (COVID-19) in March 2020 still impacts our daily lives. The detection of COVID-19 cases in addition to the safety measures –washing hands, wearing face masks, and social distancing– have proven effective to control the spread of the virus. The current detection of COVID-19 is performed with medical diagnostic tools, which are expensive, time-consuming, and generate a large amount of waste. To ease the diagnosis, we envision the use of digital health solutions powered with *Artificial Intelligence* (AI) to develop large-scale and cost-effective pre-screening tools.

To support the compliance with the safety measures, researchers have investigated the detection of face masks using visual [1] and acoustic [2] information, and the recognition of washing hands exploiting the measurements read with the sensors embedded in a smartwatch [3], [4]. AI-based solutions have also been proposed in the literature to detect COVID-19 patients from the analysis of X-ray images [5], [6], CT scans [7], [8], or acoustic signals produced by the human body, including coughs, breaths, and speech [9], [10], [11], [12], [13], [14].

Herein, we present the *Your Voice Counts* dataset, a new dataset collected using smartphones to detect COVID-19 from different vocal sounds, including the sustained vowels /a:/, /e:/, /i:/, /o:/, and /u:/, coughing, and reading samples [15]. Furthermore, we report our initial experiments conducted in this datasetset using *Convolutional Neural Networks* (CNNs) to extract deep learnt representations from the Mel-spectrogram representations of the acoustic samples. The learning of the embedded representations in traditional supervised learning approaches using CNNs for feature extraction is not constrained during the training process. This casts doubts on the meaning of the embedded representations. To overcome this issue, we propose the use of a triplet loss function to guide the learning of these embedded representations. This technique aims to minimise the distance of the embedded representations learnt from the Mel-spectrogram representations corresponding to speakers with the same COVID-19 status, and maximise the distance of the embedded representations when the Mel-spectrogram representations correspond to speakers with the opposite COVID-19 status.

The rest of the paper is organised as follows. Section II introduces the dataset explored in this work. Section III describes the methodology followed, and Section IV reports and analyses the results obtained from the experiments conducted. Finally, Section V concludes the paper.

II. YOUR VOICE COUNTS DATASET

This work explores the *Your Voice Counts* dataset, a new dataset for COVID-19 detection from vocal system-produced sounds collected in-the-wild from the general public. The study procedures were approved by the ethics representative of the University of Augsburg, Germany, and all participants gave their written informed consent for participation. For the present study, we include only German-speaking participants living in Germany or Austria to minimise language-dependent influences. The COVID-19 positive group comprises 8 participants (7 males, 1 female; mean age = 36 years ± 16 years standard deviation, age range = 17 – 59 years), while the COVID-19 negative group includes 75 participants (27 males, 48 females; mean age = 45 years ± 14 years standard deviation, age range = 18 – 78 years). All participants of the COVID-19 positive group were tested positive for COVID-19 within the last 3 days prior to inclusion into the study; all participants of the COVID-19 negative group were tested negative for COVID-19 within the last 3 days

---

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE). It has also been supported by the Bavarian Ministry of Science and Arts through the ForDigitHealth project, funded by the Bavarian Research Association on Healthy Use of Digital Technologies and Media.

[1] A. Mallol-Ragolta, F. B. Pokorny, K. D. Bartl-Pokorny, A. Semertzidou, and B. W. Schuller are with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany.

[2] F. B. Pokorny and K. D. Bartl-Pokorny are also with the Division of Phoniatrics and the Division of Physiology, Medical University of Graz, Graz, Austria.

[3] B. W. Schuller is also with the Group on Language, Audio, & Music, Imperial College London, London, UK.

* Corresponding author: `adria.mallol-ragolta@uni-a.de`

TABLE I

Summary of the number of German speakers of the Your Voice Counts dataset selected to populate the 3 folds in which we have partitioned the available data. The partitioning is performed following a stratification approach along the dimensions of COVID-19 status, sex, and age.

		Fold 1			Fold 2			Fold 3			$\sum$
COVID-19 Status		Pos	Neg	$\sum$	Pos	Neg	$\sum$	Pos	Neg	$\sum$	
Age $\leq$ 30	M	1	–	1	1	1	2	1	1	2	5
	F	–	4	4	–	4	4	1	3	4	12
	$\sum$	1	4	5	1	5	6	2	4	6	17
30 $<$ Age $<$ 60	M	–	6	6	2	6	8	2	6	8	22
	F	–	11	11	–	11	11	–	9	9	31
	$\sum$	–	17	17	2	17	19	2	15	17	53
Age $\geq$ 60	M	–	1	1	–	3	3	–	3	3	7
	F	–	2	2	–	2	2	–	2	2	6
	$\sum$	–	3	3	–	5	5	–	5	5	13
$\sum$		1	24	25	3	27	30	4	24	28	83

TABLE II

Summary of the vocal sounds included in the Your Voice Counts dataset, and the amount of available data in terms of time per fold. The temporal information is provided in the format (HH:)MM:SS.

Vocal Sound	Fold 1	Fold 2	Fold 3	$\sum$
/a:/	4:42	5:50	6:30	17:02
/e:/	5:00	6:02	6:49	17:51
/i:/	5:06	5:48	6:45	17:39
/o:/	5:21	6:06	6:43	18:10
/u:/	4:45	5:59	6:51	17:35
Coughing	6:03	7:52	8:22	22:17
Reading	20:39	25:45	23:53	1:10:17
$\sum$	51:36	1:03:22	1:05:53	3:00:51

prior to inclusion into the study. The participants provided a copy of their COVID-19 test result and completed a short questionnaire including information about potential symptoms and pre-existing health issues. 2/8 COVID-19 positive participants were smokers, 0/8 reported pre-existing pulmonary diseases or voice problems, and 8/8 reported fever and/or respiratory symptoms at the time of recording. 4/75 COVID-19 negative participants were smokers, 7/75 reported pre-existing pulmonary diseases or voice problems, and 23/75 reported fever and/or respiratory symptoms at the time of recording. None of the participants reported reading problems.

According to the study protocol, the speakers were asked to record themselves when performing the following three tasks. Task 1: Production of the sustained vowels /a:/, /e:/, /i:/, /o:/, /u:/ (in this order), representing phonemes of the German standard language. The participants were instructed to produce each vowel as long as possible and to make a breathing break after each vowel. Task 2: The participants were asked to cough deliberately 5 times and to make a breathing break after each cough. Task 3: Reading aloud the standard phonetic text "The North Wind and the Sun" in German. From the technical perspective, the participants were instructed to take the recordings in a quiet room using their own smartphone at a distance of approximately 40 centimetres from the face. Each participant transferred his/her recording via the secure file-sharing service of the University of Augsburg. Upon file receipt, the recordings were converted into the audio format 16 kHz/16 bit single-channel PCM and segmented manually for the different tasks.

The selected participants are split into 3 stratified folds to train and assess the performance of our COVID-19 detection models. The stratification is performed along three dimensions: COVID-19 status (positive or negative), sex (male or female), and age (below 30, between 30 and 60, and above 60). Each fold contains both COVID-19 negative participants with symptoms and/or pre-existing health issues, and healthy COVID-19 negative participants. Table I summarises the number of participants belonging to each fold in terms of these three dimensions. The total amount of data time-wise of the different vocal sounds included in the dataset per fold is reported in Table II.

III. METHODOLOGY

This section describes the methodology followed. Section III-A details the pre-processing applied to the recorded vocal sounds, Section III-B introduces the models implemented, and Section III-C summarises their training details.

A. Data Preparation

This study analyses the 7 vocal sounds available in the dataset (cf. Section II) separately. Nonetheless, the same pre-processing is applied to all of them. To guarantee that the networks can get enough information to extract embedded representations from, we require each vocal sound to have a minimum length of 5 sec. Shorter vocal sounds are extended via replication until reaching this threshold. From each vocal sound, we extract the Mel-spectrogram representation in dBs, using 128 *Mel-Frequency Cepstral Coefficients* (MFCCs), and a hop size of 128 samples (8 msec). The obtained representation is normalised so its values range $\in [0, 1]$. After the normalisation, we segment the generated Mel-spectrogram using a window length of 5 sec –625 bins in the temporal domain– and a 50 % overlap. The segmented representations are stored as images of 224×224 pixels.

B. Models Description

The models implemented receive a Mel-spectrogram representation of 5 sec length as input, and output the probability of the current sample to belong to a COVID-19 positive or negative speaker. The networks behind are composed of two blocks: the first block extracts deep learnt representations from the input Mel-spectrogram representations, while the second block performs the actual classification.

The feature extraction block implements two convolutional layers with 32 and 64 filters, respectively, with a kernel size of 3×3, and a stride of 1. Following each convolutional layer, we use batch normalisation, and the output is transformed using a *Rectified Linear Unit* (ReLU) function. A 2-dimensional max pooling layer, and a 2-dimensional adaptive average pooling layer are included at the end of the first and second convolutional blocks, respectively. The output of the feature extraction block is flattened before being fed into the classification block of the network, generating a 256-dimensional embedded representation of the input Mel-spectrogram representation. The classification block contains two fully connected layers with 32 and 2 output neurons, respectively, preceded each by a dropout layer with probability 0.3. While the output of the first layer is transformed using a ReLU activation function, the output of the second layer is transformed using a Softmax function, so the network outputs can be interpreted as probability scores. When multiple Mel-spectrogram representations are generated from the same vocal sound, we use a majority voting schema to determine the COVID-19 status to infer for the overall sample.

The baseline models implementing the aforementioned network architecture are trained following a traditional supervised learning approach –i. e., the network parameters are updated by feeding a sample into the model, comparing the output and the ground truth information, and back-propagating the error. In an attempt to improve the quality of the embedded representations learnt, we explore the use of a triplet loss-based approach. When training the models using this approach, we feed into Siamese feature extraction blocks a Mel-spectrogram representation of the current speaker, the anchor X^a, accompanied by 2 additional, randomly selected representations from the training data: one of these representations belongs to a speaker with the same COVID-19 status as the anchor, X^+, and the other belongs to a speaker with the opposite COVID-19 status as the anchor, X^-. We represent the embedded representations learnt as $f(X^a)$, $f(X^+)$, and $f(X^-)$, respectively. In the next step, we compute the Euclidean distance between the embedded representations of X^+ and X^-, and the anchor; i. e.,

$$d^+ = || f\left(X^+\right) - f\left(X^a\right) ||_2^2, \tag{1}$$

and

$$d^- = || f\left(X^a\right) - f\left(X^-\right) ||_2^2, \tag{2}$$

respectively. Using these distances, we define the triplet loss

$$\mathcal{L} = max\{0, d^+ + \alpha - d^-\}. \tag{3}$$

Updating the network parameters of the feature extraction block using this loss function, we aim to improve the robustness of the embedded features learnt increasing the separability of the inter-class embedded representations. Finally, only the embedded representation learnt from the anchor representation, $f(X^a)$, is fed into the classification block of the network.

C. Networks Training

For a fair comparison of the models, these are all trained under the exact same conditions. The baseline models use the Categorical Cross-Entropy as the loss to minimise. The triplet loss-based models update the network parameters of the feature extraction block by minimising the loss function defined in Equation (3) using $\alpha = .5$, which we set empirically, and the network parameters of the classification block by minimising the Categorical Cross-Entropy loss. For the optimisation, we use Adam with a fixed learning rate of 10^{-3}. To compensate for the data imbalance in terms of the COVID-19 status (cf. Table I), we implement a weighted random sampler to select the Mel-spectrograms for training at each epoch. We assess the model performances using a nested 3-fold cross-validation. We select the *Unweighted Average Recall* (UAR) as the evaluation metric, and, therefore, we define $\mathcal{L}_{UAR} = 1 - UAR$ as the validation error to monitor during the training process. Network parameters are updated in batches of 128 samples, and trained during a maximum of 100 epochs. We implement an early-stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. Each fold is trained during a specific number of epochs. Hence, when modelling all training material and to prevent overfitting, the training epochs are determined by computing the mean of the training epochs processed in each fold, rounded up to the next integer.

IV. EXPERIMENTAL RESULTS

The results obtained with the baseline and the triplet loss-based models are presented in Table III.

Analysing the average results obtained with the baseline models (cf. Table III) among the 3 folds, we observe that the models perform worse than chance level (50 % for two classes using UAR) when exploiting the /u:/ and the coughing sounds, with a UAR of 49.2 % and 41.2 %, respectively. The baseline models only achieve a chance level performance with the /a:/ and the /i:/ sounds. This result suggests that, using these sounds, the baseline models most likely predict that all samples belong to the same class. Using the /e:/, the /o:/, and the reading sounds, the baseline models perform better than chance, scoring a UAR of 57.4 %, 54.4 %, and 51.8 %, respectively.

The results obtained with the triplet loss-based models (cf. Table III) indicate their underperformance when exploiting

TABLE III

DESCRIPTIVE STATISTICS (μ: MEAN, σ: STANDARD DEVIATION) COMPUTED FROM THE UNWEIGHTED AVERAGE RECALL SCORES (%) OBTAINED AMONG THE 3 FOLDS IN WHICH WE PARTITIONED THE AVAILABLE DATA PER VOCAL SOUND USING THE BASELINE AND THE TRIPLET LOSS-BASED MODELS. THE CHANCE LEVEL FOR A BINARY CLASSIFICATION PROBLEM IN TERMS OF UAR IS 50 %.

Vocal Sound	Baseline		Triplet Loss	
	μ	σ	μ	σ
/a:/	**50.0**	0.0	42.8	6.2
/e:/	57.4	12.8	**59.2**	17.7
/i:/	50.0	0.0	**65.1**	19.3
/o:/	**54.4**	14.3	53.5	6.0
/u:/	49.2	5.0	**56.9**	12.0
Coughing	41.2	11.8	**50.6**	1.1
Reading	51.8	5.0	**66.5**	21.1

the /a:/ sounds. This model scores a UAR of 42.8 %, below the chance level. The triplet loss-based model using participants' voice while reading achieves the best performance in terms of UAR, 66.5 %, closely followed by the model exploiting the /i:/ sounds with a UAR of 65.1 %.

Comparing the performances obtained with the baseline and the triplet loss-based models, we observe that the triplet loss-based models outperform the baseline models in 5 out of 7 vocal sounds investigated. Thus, it seems reasonable to state that the triplet loss-based approach contributes to learn more discriminative inter-class embedded representations, which help to improve the model performances. Among all the experiments conducted, the triplet loss-based model exploiting the reading sounds scores the best UAR of 66.5 %.

V. CONCLUSIONS

This work presented the *Your Voice Counts* dataset, a new dataset for detecting COVID-19 from vocal sounds. The dataset included German speakers and was recorded using smartphones. Furthermore, we focused on the use of a triplet loss function to train COVID-19 detection models from the collected vocal sounds. With the use of the triplet loss function, we aimed at guiding the training of the deep learnt representations towards more discriminative and robust inter-class representations. The results obtained from the experiments conducted supported the suitability of the triplet loss function, as the models trained using this approach outperformed the baseline models in 5 out of the 7 vocal sound types investigated. The best UAR score of 66.5 % was obtained using the triplet loss-based model exploiting reading sounds. As observed from Table II, the training material available from this sound type was larger than from the sustained vowels, or the coughing sounds. Thus, this attribute can benefit the models using the reading sounds. We hypothesise that the suitability of this vocal sound for the detection of COVID-19 could also be attributed to the tasks order in the recording protocol. The reading task was the last one, and, as a consequence, the vocal system was stressed longer, increasing the salient information of the COVID-19 positive participants. Future works can consider the fusion of different vocal sounds, the exploration of more advanced techniques to increase the separability of the inter-class representations, and the use of few-shot learning to overcome the scarcity of COVID-19 positive samples.

REFERENCES

[1] S. Rashmi Nayak and N. Manohar, "Computer-Vision based Face Mask Detection using CNN," in *Proceedings of the 6th International Conference on Communication and Electronics Systems*, Coimbatre, India, 2021, pp. 1780–1786, IEEE.

[2] 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, 2021, pp. 2079–2082, IEEE.

[3] A. Mallol-Ragolta, A. Semertzidou, M. Pateraki, and B. Schuller, "harAGE: A Novel Multimodal Smartwatch-based Dataset for Human Activity Recognition," in *Proceedings of the 16th International Conference on Automatic Face and Gesture Recognition*, Jodhpur, India – Virtual Event, 2021, IEEE, 7 pages.

[4] A. Mallol-Ragolta, A. Semertzidou, M. Pateraki, and B. Schuller, "Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition," *Frontiers in Computer Science, section Mobile and Ubiquitous Computing*, vol. 4, pp. 1–10, 2022, Article ID 796866.

[5] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Systems with Applications*, vol. 164, 2021, Article ID: 114054, 11 pages.

[6] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, vol. 24, pp. 1207–1220, 2021.

[7] P. Gifani, A. Shalbaf, and M. Vafaeezadeh, "Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 115–123, 2021.

[8] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning," *Sensors*, vol. 21, no. 2, 2021, Article ID: 455, 22 pages.

[9] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," in *Proceedings of ICASSP*, Toronto, ON, Canada, 2021, pp. 8328–8332, IEEE.

[10] T. K. Dash, S. Mishra, G. Panda, and S. C. Satapathy, "Detection of COVID-19 from speech signal using bio-inspired based cepstral features," *Pattern Recognition*, vol. 117, 2021, Article ID: 107999, 13 pages.

[11] A. Mallol-Ragolta, H. Cuesta, E. Gómez, and B. Schuller, "Cough-based COVID-19 Detection with Contextual Attention Convolutional Neural Networks and Gender Information," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 941–945, ISCA.

[12] S. Liu, A. Mallol-Ragolta, and B. Schuller, "COVID-19 Detection with a Novel Multi-Type Deep Fusion Method using Breathing and Coughing Information," in *Proceedings of the 43rd Annual International Conference of the Engineering in Medicine & Biology Society*, Guadalajara, Mexico – Virtual Event, 2021, pp. 1840–1843, IEEE.

[13] P. Hecker, F. Pokorny, K. Bartl-Pokorny, U. Reichel, Z. Ren, S. Hantke, F. Eyben, D. Schuller, B. Arnrich, and B. Schuller, "Speaking Corona? Human and machine recognition of COVID-19 from voice," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 1029–1033, ISCA.

[14] T. Xia, D. Spathis, C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta, and C. Mascolo, "COVID-19 Sounds: A Large-Scale Audio Dataset for Digital COVID-19 Detection," in *Proceedings of the 35th Conference on Neural Information Processing Systems – Track on Datasets and Benchmarks*, Virtual Event, 2021, NeurIPS, 13 pages.

[15] K. D Bartl-Pokorny, F. B Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. Schuller, "The voice of COVID-19: Acoustic correlates of infection in sustained vowels," *Journal of the Acoustical Society of America*, vol. 149, pp. 4377–4383, 2021.