SARS-CoV-2 Detection From Voice

Gadi Pinkas, Yarden Karny, Aviad Malachi, Galia Barkai, Gideon Bachar, and Vered Aharonson

***Abstract*—Automated voice-based detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) could facilitate the screening for COVID19. A dataset of cellular phone recordings from 88 subjects was recently collected. The dataset included vocal utterances, speech and coughs that were self-recorded by the subjects in either hospitals or isolation sites. All subjects underwent nasopharyngeal swabbing at the time of recording and were labelled as SARS-CoV-2 positives or negative controls. The present study harnessed deep machine learning and speech processing to detect the SARS-CoV-2 positives. A three-stage architecture was implemented. A self-supervised attention-based transformer generated embeddings from the audio inputs. Recurrent neural networks were used to produce specialized sub-models for the SARS-CoV-2 classification. An ensemble stacking fused the predictions of the sub-models. Pre-training, bootstrapping and regularization techniques were used to prevent overfitting. A recall of 78% and a probability of false alarm (PFA) of 41% were measured on a test set of 57 recording sessions. A leave-one-speaker-out cross validation on 292 recording sessions yielded a recall of 78% and a PFA of 30%. These preliminary results imply a feasibility for COVID19 screening using voice.**

***Index Terms*—COVID19, audio embeddings, transformer, recurrent neural network, ensemble stacking, semi supervised learning.**

***Impact Statement*—A deep machine learning model was trained and tested on audio recordings from subjects who underwent SARS-CoV-2 testing at the time of recording. The strength of self-supervised pre-training, regularization techniques and ensemble stacking were demonstrated. The model indicates a feasibility for COVID19 screening using self-recorded voice.**

Manuscript received July 29, 2020; revised August 31, 2020 and September 11, 2020; accepted September 11, 2020. Date of publication September 24, 2020; date of current version October 14, 2020. This work was supported by the Directorate of Defense Research and Development (DDRD) of the Israeli Ministry of Defense. *(Corresponding author: Vered Aharonson.)*

Gadi Pinkas, Yarden Karny, and Aviad Malachi are with the Afeka Center of Language Processing, Afeka, Tel Aviv Academic College of Engineering, Tel Aviv-Yafo 6910717, Israel.

Galia Barkai is with the Pediatric Infectious Diseases Unit, Safra Children's Hospital, Sheba Medical Center and Sackler School of Medicine, Tel-Aviv University, Tel Aviv-Yafo 69978, Israel.

Gideon Bachar is with the Department of Otorhinolaryngology, Rabin Medical center and Sackler School of Medicine, Tel-Aviv University, Tel Aviv-Yafo 69978, Israel.

Vered Aharonson is with the Afeka Center of Language Processing, Afeka, Tel Aviv Academic College of Engineering, Tel Aviv-Yafo 6910717, Israel, with the School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa (e-mail: vered.aharonson@wits.ac.za).

Digital Object Identifier 10.1109/OJEMB.2020.3026468

I. Introduction

A RELIABLE detection of COVID19 through audio processing of speech, cough and breathing could facilitate a globally accessible screening. Automated audio processing of sounds associated with respiratory diseases traditionally employed feature extraction and classifiers, convolutional neural networks (CNN) and recurrent neural networks (RNN). Two recent reviews on breathing, cough and speech analysis reported detection or classification of sounds and symptomatic vocal attributes which were associated with respiratory diseases. The accuracy values reported were between 70% and 100% [1]–[3]. The diseases studied were obstructive sleep apnea and Tuberculosis, asthma, bronchitis and pneumonia. These acute diseases are manifested by symptoms such as nasal obstruction, cough and fatigue, which are detectable in vocal sounds. Although these symptoms are prevalent in COVID19, screening for early stages of the disease where symptoms may be milder and accompanied by a variety of other symptoms is more challenging [4]–[6]. A recent study on COVID19 speech data used feature extraction and a classifier. A classification of three levels of COVID19 symptoms severity yielded an accuracy of 0.65–0.66 [7]. No control data was used in this recent study, however, making it impractical for COVID19 screening.

The subtle and mixed symptoms of early COVID19 and their uncertain attributes in patients' voice, promote a usage of deep neural networks (DNNs) for COVID19 screening. Specifically, RNNs are used for handling variable-length input sequences [8]. This property is needed since prevalent COVID19 symptoms like fatigue may stretch and change over a large and varying number of frames in the recorded voice [6], [7].

Long sequences, however, may produce diminishing or exploding gradients in RNNs, even in those designed for long- and short-term memorization [8], [9]. Attention mechanisms in DNNs learn to focus on areas within long and variable-length sequences [10], [11]. In an implementation of multiple attention heads, each head learns to focus on different areas of the sequence and relates to these areas in the subsequent classification layers [10], [11]. Transformer-based DNNs include attention mechanisms that indicated better performance compared to RNNs in vision, language translation, speech recognition and sentiment analysis [12]–[16], [18]. The transformer-based architectures can be used in a self-learning setting where they are pre-trained on unlabeled audio recordings. These systems were able to automatically discover more useful features compared to the standard Mel-frequency features in a variety of audio applications [13], [14], [17].

The challenging voice-based COVID19 detection may benefit from an "ensemble of experts" stacking: a collection of high-variance classifiers (expert sub-models) combined in a meta-model [19]–[21]. In this paradigm, each expert classifier specializes in a specific input type. In the context of COVID19 detection these input types may be speech, cough or breathing. All "expert opinions" are then weighed and combined in a meta-classifier.

An additional challenge in deep learning for voice-based COVID19 screening is that available datasets are either small or unlabeled. Evidence from COVID19 literature, as well as discussions with healthcare professionals attending COVID19 patients convey that different patterns may be detected within different types of vocal utterance [7]. A large set of voice inputs, and hence long recording sessions, are therefore needed to enable a rigorous search for these patterns. Reliable labeling of Sars-Cov2 positive and negative need to be performed. This necessitates the use of ground truth examinations concurrently with the recording. The complexity and costs involved in this data collection thus yield small datasets. Deep models tend to have low bias when fitted to highly non-linear data, but can easily over-fit to small datasets and generate high variance error.

Semi-supervised learning combines unsupervised pre-training of DNNs on large unlabeled dataset of recordings, followed by fine-tuning of the pre-trained DNN using smaller, labelled dataset. A transformer is an implementation of this method which was found effective in discovering useful audio features and in reducing overfitting [13], [14], [17]. This paradigm can be applied in voice-based screening of COVID19: The transformer self-learns to transform sequences of audio frames into fixed length vector representations called embeddings. These embeddings may be viewed as features that the transformer learned to extract automatically from variable length recordings, and which can be used to classify labelled COVID19 patients' voices.

Augmentation and regularization techniques are useful for deep learning in small datasets. Prevalent augmentation methods, however, add noise or change frequency-components and may corrupt the subtle COVID19 patterns. Bootstrapped sampling techniques, that merely crop random segments of the original recordings, can provide augmentation without inducing noise or distorting the spectrum. In addition, bootstrapped sampling produces shorter audio sequences and thus reduces memory consumption and training time [22]. Regularization techniques provide a complementary solution to DNN overfitting [23].

The current study applied an architecture that combines the abovementioned techniques. The system was trained and tested on recordings acquired from a cohort of verified Sars-CoV-2 positive and negative subjects. The focus of the study is the Sars-CoV-2 detection system and its performance. The data acquisition experiment and protocol are briefly described in the next section.

II. Materials and Methods

A. Dataset

This study made use of a dataset collected in a national COVID19 data collection project, led by the Israeli Directorate of Defense Research and Development. The dataset, here forth denoted "COVID19 dataset" included recordings of 29 SARS-CoV-2 positive and 59 negative controls. All subjects were tested for SARS-CoV-2 using Real-Time Polymerase Chain Reaction (RT-PCR) at the time of recording. Age and gender data were collected by the healthcare team who recruited the subjects. Ethics approvals were obtained for each testing center. All subjects signed an informed consent, downloaded the recording application on their smartphones and were instructed to follow the guidance and prompts provided by the application. The application was developed by Vocalis according to a data acquisition protocol designed by a multi-disciplinary team of speech processing scientists, speech therapists and healthcare professionals treating COVID19 patients. The vocal inputs in the recordings were the phonemes /ah/ and /z/, coughs and counting from 50 to 80. The phoneme /ah/ is traditionally used in speech-based diagnosis, as a representative vowel sound [24]. The phoneme /z/ is a representative sonorous consonant for voiced (involving the vocal chords) sound [25], [26]. Counting is a form of continuous speech that does not require cognition and which is clean of social or emotional bias. In the respiratory diseases context the length of the counting induce vocal fatigue [27]. The cough is a non-speech sound that is symptomatic in respiratory disease [3]. The set of these four input types thus contained short speech utterances -phonemes, long speech utterance – counting and non-speech voicing – the cough.

The dataset recording format was WAV at a sampling frequency of 32 kHz. The recording setup was kept constant for all subjects. The subjects recorded in their rooms, either at the hospitals or in isolation sites. The application prompted the subjects to sit down and to place the smartphone on a table, at a distance of 20 cm from their mouth. The application then prompted the subjects to repeat each phoneme three times and to produce an acted cough three times. The counting prompt instructed the subjects to "take a deep breath and count from 50 to 80 as fast as you can". The recording was terminated after 10 seconds for each phoneme and the cough, and after 50 seconds for the counting. The subjects were asked to record themselves daily for 14 days.

The recordings were divided into a training set of 70 speakers and a test set of 18 speakers. Sixty of the 88 subjects, both SARS-CoV-2 positives and SARS-CoV-2 negatives, repeated their recording during 2 up to 14 days. Repeated recordings were considered as separate events, yielding 235 events for the training set and 57 events for the test set. The training and test sets were balanced to include similar distribution of events across the subjects in the sets. SARS-CoV-2 positive and negative classes, age and gender were balanced between the training and test sets.

B. High-level Machine Learning Architecture

The architecture supported a semi-supervised learning approach and consisted of three stages.

At stage 1 the transformer of [13] was pre-trained on the unlabeled recordings of the Librispeech dataset. The self-training's objective was to reconstruct missing frames in the recordings

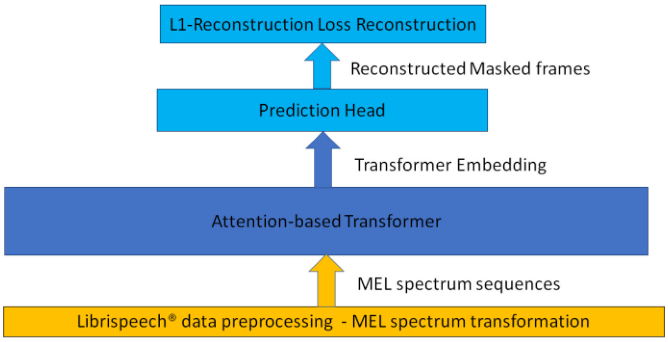
L1-Reconstruction Loss Reconstruction
Reconstructed Masked frames
Prediction Head
Transformer Embedding
Attention-based Transformer
MEL spectrum sequences
Librispeech® data preprocessing - MEL spectrum transformation

Fig. 1. A block diagram of stage 1: Pre-training of a transformer. Librispeech unlabeled recordings pre-processing (orange block) produced MEL-spectrum frame sequences. The self-trained transformer (dark blue block) learned to transform the frame sequences into "Transformer Embeddings" which were later used in stage 2. During pre-training only, a prediction head (light blue block) was used to reconstruct the missing frames using L1 loss.

[13]. The values generated in the last layer of the transformer were the embeddings that were used as features at stage 2.

At stage 2 the COVID19 dataset was used to train RNN-based expert classifiers [17]. Each expert specialized in one or multiple types of vocal input (/z/, /ah/, counting or cough). For this vocal input, the classifier produced a score reflecting its "expert opinion" on the probability of the speaker to be Sars-CoV-2 positive. The small COVID19 dataset was segmented using bootstrapped sampling [22] and was transformed into sequences of Mel-spectrum representation. Embeddings were generated for each pre-processed sample using the pre-trained transformer of stage 1. Each expert classifier used the embeddings to produce a score - Sars-CoV-2 positive probability - per sample.

At stage 3, the scores of each expert classifier were averaged across samples per speaker. These average scores, from all the experts, were assembled into a feature vector per speaker. The feature vectors were used to train a linear support vector machine (SVM) that weighed the scores of each expert and predicted the final score.

The three stages are illustrated in the block diagrams of Figures 1, 2 and 3 and their implementation details are described in the next section.

C. Audio Processing and Machine Learning Implementation

Stage 1, illustrated by the block diagram in Figure 1, consisted of pre-processing and transformer pre-training. The pre-processing (orange block in Figure 1) converted Librispeech unlabeled recordings into Mel-spectrum representation using Librosa [11], [28]. This produced 25ms frames with 10ms overlap, where each frame consisted of 80 MEL-scaled frequencies and 80 first derivatives.

The transformer (blue block in Figure 1) used the basic setting of the Mockingjay system [13]. This architecture included sinusoidal positional encoding [12] and a stack of three double layers. Each double layer consisted of a sublayer made of 12 attention heads [12] and a fully connected sub-layer. Residual connections [30] and layer normalization [31] were implemented between the sub-layers. An additional two-layer prediction head [13] was used only for the pre-training (light blue block in Figure 1), and was not used in stage 2. The transformer performed self-training by randomly masking the input frames while using L1 loss to reconstruct the missing frames [13]. The values of the transformer's last layer were taken as the embedded representations of the input samples and were used as input features in stage 2.

In stage 2, illustrated in Figure 2, expert sub-models were trained to classify the COVID19 dataset recordings. The expert classifiers in this stage were trained either on one input type, i.e. a phoneme, counting or cough, or on all input types, the latter becoming a "multi-disciplinary" expert classifier.

The pre-processing (orange block in Figure 1) included down-sampling to 16 kHz and bootstrapped sampling by randomly segmenting each recorded file 20 times. The starting point of

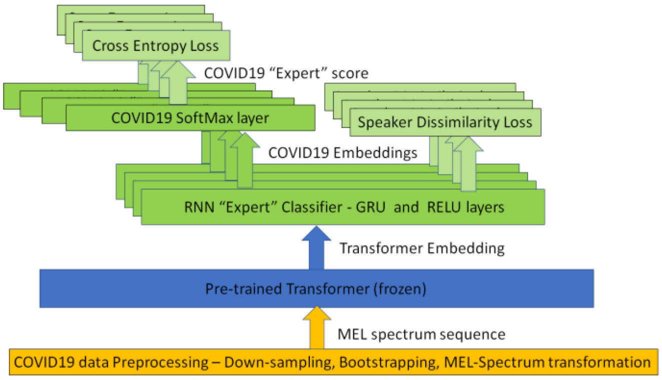
Cross Entropy Loss
COVID19 "Expert" score
COVID19 SoftMax layer
Speaker Dissimilarity Loss
COVID19 Embeddings
RNN "Expert" Classifier - GRU and RELU layers
Transformer Embedding
Pre-trained Transformer (frozen)
MEL spectrum sequence
COVID19 data Preprocessing – Down-sampling, Bootstrapping, MEL-Spectrum transformation

Fig. 2. A block diagram of stage 2: training of expert classifiers. COVID19 dataset recordings were pre-reprocessed (orange block) by bootstrapping and conversion into MEL-spectrum frame sequences. The frame sequences were then fed into the pre-trained transformer (blue block). Each expert classifier consisted of a GRU RNN that its last layer was fully connected to a RELU layer and then to a SoftMax layer (green blocks). The classifier was trained on the transformer embeddings and predicted an "expert opinion" score—The estimated probability of Sars-COV-2 positive. Cross-entropy loss was used in the classification and speaker dissimilarity loss was added for regularization.

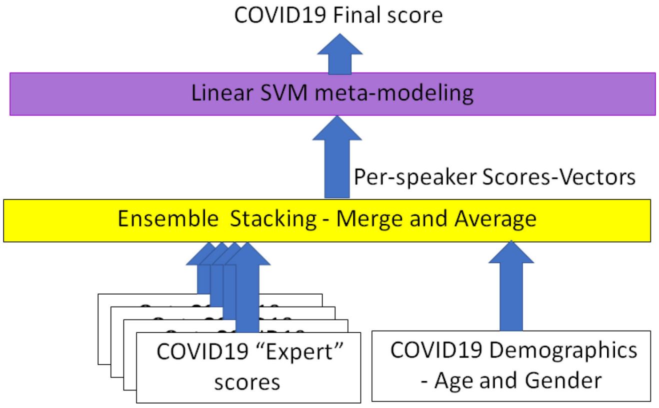
COVID19 Final score
Linear SVM meta-modeling
Per-speaker Scores-Vectors
Ensemble Stacking - Merge and Average
COVID19 "Expert" scores
COVID19 Demographics - Age and Gender

Fig. 3. A block diagram of stage 3: Training a meta-model using ensemble stacking. The averaged sample-scores of the best experts augmented by gender and age were stacked in an Ensemble by merging and averaging them per speaker (yellow block) to create Scores-Vectors. A meta-model (purple block) is trained using a linear SVM to predict the final score.

each segment was taken with replacement from a uniform distribution. The sample length was taken from a normal distribution with a mean of 1.5s and a standard deviation of 0.1s. The lengths were truncated at ±3 standard deviations from the mean.

The bootstrapped samples were transformed into Mel- spectrum frame sequences, as in the preprocessing of stage 1. The frame sequences were fed into the pre-trained transformer created in stage 1 (blue block of Figure 2), where the transformer weights were kept frozen. The embeddings generated by the transformer were the inputs for the expert classifiers (green blocks in Figure 2). The embeddings were aggregated using an RNN of 768 gated recurrent units (GRU) [29]. The last layer of the RNN was fed into a fully connected layer of 100 Rectified Linear Units (RELU) [32]. The RELU layer produced vector values, denoted "COVID19 embeddings". These values were the high-level features which could hypothetically discriminate audio samples of Sars-CoV2 positives from negatives. A SoftMax layer computed the probabilities for SARS-CoV-2 positive. The training used the standard Cross Entropy loss.

The techniques employed in order to avoid overfitting in the training of the classifiers in stage 2 were: a) aggressive dropout [34] of 60% to 70%, b) weight decay rate of 0.1, c) early termination, when the estimated balanced accuracy measured on the validation set stopped dropping, and d) a novel speaker-dissimilarity regularization (SDR) loss. The SRD loss was computed on the COVID19 embeddings (light green blocks in Figure 2). The purpose of the SDR (eq. 1) was to generate speaker-invariant embeddings by penalizing the average of L2 dissimilarities of the COVID19 embeddings across different speakers of the same class (Sars-CoV-2 positive or negative) within each training batch:

$$SDR(batch) = \alpha \cdot average_{i<j}(\|v_i - v_j\|_2)$$
$$v_i, v_j \in batch\ and\ t\,(v_i) = t\,(v_j)\,,$$

where t(ν) was the class of embedding ν, and α was a hyper-parameter denoting the SDR penalty assigned to dissimilarities within each class. During training, the SDR loss was added to the cross-entropy loss and the sum was minimized using Adam optimizer [33].

To select the best sub-models, each expert was trained with 5 different seeds while varying the drop-out, the RELU layer size, the learning rate, the weight decay rate and the SDR penalty α. A 5-fold speaker cross validation was used to optimize the network's hyper-parameters. The best sub-models for each expertise were selected based on the F1 measure of the cross validation. The 5-fold cross validation was implemented by taking out for validation 20% of the speakers– with all their recording samples – five times.

Stage 3 (Figure 3) implemented a meta-model training that weighed each of the sub models - "expert opinions" - from stage 2 into a final score - probability of SARS-CoV-2 positive. An ensemble stacking (yellow block in Figure 3) merged the "expert opinion" scores of each expert across the 5 validation sets of stage 2, per sample. The scores of each expert were then averaged over samples, per speaker. The averaging operation was employed in order to smooth the effect of outliers, e.g. occasional samples of silence. The merge and average process yielded a vector containing the average scores of all experts for each speaker. The age and gender of the speaker were added to this vector. The resulting per-speaker scores-vectors were used by a linear SVM (purple block in Figure 3) to train a meta-model. A regularization parameter C = 1 was used for the SVM.

The SVM provided predictions on the test-set. A leave one speaker out (LOSO) cross validation, where the SVM predictions were computed on the left-out speakers, provided the validation scores. To increase robustness, the LOSO training set included the combined training and test datasets.

D. Performance Evaluation

An ablation analysis was performed in order to compare the proposed architecture with a simple baseline and to evaluate the contribution of components in the architecture.

Five experiments were performed and compared, as listed in Table I. Experiment 0 was an analogue to a linear classifier using traditional audio features. This experiment served as a baseline for the deeper networks' performance. Experiments 1 through 3 increased non-linearity by gradually adding more layers from the expert classifiers architecture. Experiment 4 implemented the full architecture and provided the meta-model final scores. Experiment 0 and 1 enabled only fixed size inputs. The deeper architecture in experiments 2, 3 and 4 allowed for longer and variable-length inputs.

The sub-models' results were evaluated using F1, Recall (sensitivity) and Probability of False Alarm (PFA). A receiver operation curve (ROC) was calculated to illustrate the performance of the meta-model (experiment 4). The ablation experiments were compared using the F1 measure.

III. Results

Seven sub-models were selected at stage 2 (experiment 3 in Table I). Six were single-input experts, trained on a single input type, and one was a "multi-disciplinary" expert, trained on all four input types. The performance metrics of these seven sub models on the test set are presented in Table II. The dropout parameter for each selected expert is given in the Table. The optimal training parameters for all seven sub models were a learning rate of 0.4E-4, a weight decay of 0.1, a RELU layer size of 100 and $\alpha = 0.1$.

Table III presents the performance metrics for the meta-model of stage 3 (experiment 4 in Table I). Both the test set and the LOSO cross-validation performance are displayed. The results are illustrated by the receiver operating curves (ROCs) in figures 4 and 5, for the test set and the LOSO cross validation, respectively. The areas under curve were 0.63 for the test set and 0.81 for the LOSO cross validation. The two ROCs differ primarily in the low PFA region: The recall values in the test set ROC are lower than 0.5 for PFA values of 0.30 and lower. The recall values in the LOSO cross-validation ROC steadily increase in this PFA region from 0.5 to 0.79 for the PFA range 0.1 to 0.30.

Table IV presents an ablation analysis of experiments 0 to 3 in Table I. F1 performance measures for the four input types are presented. The results of experiment 3, where the pre-trained

TABLE I
EXPERIMENTS FOR PERFORMANCE EVALUATION

Experiment	Description
0	A baseline: Fixed sized blocks of 50 input frames were directly connected to the SoftMax output layer (similar to logistic regression)
1	A 100-RelU layer was added between the fixed sized inputs and the SoftMax output layer.
2	A 768 GRUs RNN layer was added between the inputs and the RELU layer, allowing inputs samples of variable length.
3	An addition of the pre-trained transformer (fig. 2).
4	A full system of transformer and ensemble stacking

TABLE II
TEST RESULTS OF THE BEST PERFORMING SUB MODELS

Expertise	Dropout	Recall	PFA	F1	Precision
/z/	0.6	0.80	0.45	0.81	0.82
counting	0.6	0.80	0.56	0.80	0.80
[/z/, /ah/, cough, counting]	0.7	0.78	0.50	0.79	0.80
/z/	0.7	0.76	0.42	0.79	0.82
/ah/	0.6	0.66	0.34	0.74	0.83
counting	0.7	0.61	0.30	0.71	0.85
cough	0.6	0.48	0.39	0.58	0.72

TABLE III
PERFORMANCE OF THE ENSEMBLE STACKING META-MODEL

Dataset	Recall	PFA	F1	Precision
Test	0.78	0.41	0.79	0.79
Cross-Validation	0.78	0.30	0.74	0.71

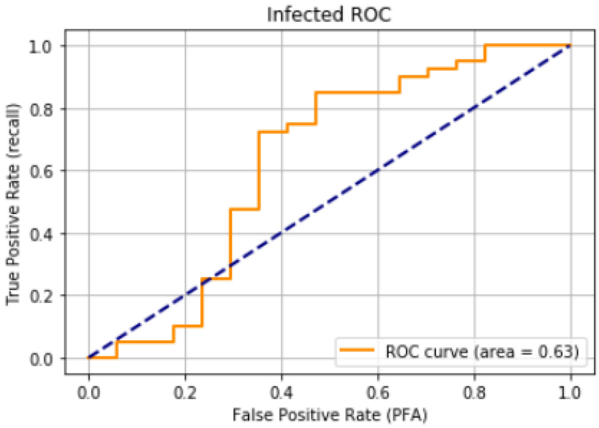

Fig. 4. Receiver Operating Curve (ROC) of the ensemble stacking on the test set.

TABLE IV
F1 PERFORMANCE IN THE ABLATION EXPERIMENTS

Input	Ex0	Ex1	Ex2	Ex3
/z/	0.70	0.70	0.73	0.81
/ah/	0.69	0.69	0.71	0.74
cough	0.58	0.43	0.41	0.58
counting	0.67	0.64	0.71	0.80

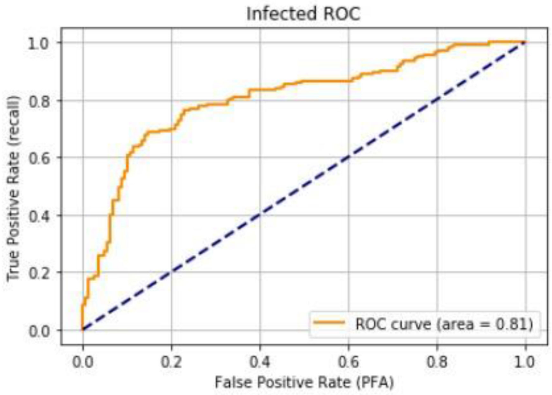

Fig. 5. Receiver Operating Curve (ROC) of the ensemble stacking, using Leave one speaker out cross validation.

transformer was used, reflect the results in Table II. The F1 value in this experiment was higher compared to the baseline provided by experiment 0, where Mel-frequency features were directly fed into a SoftMax layer. Higher F1 measures are demonstrated for all inputs except for the cough, where the F1 value was similar in both experiments. The counting input yielded the highest increase in F1, from 0.67 to 0.80. Experiments 1, where an intermediate RELU hidden layer was added between to the input and the SoftMax layer, reflects similar F1 values for the /z/ and /ah/ inputs and lower values for the counting and cough inputs, compared to the baseline experiment. Experiments 2, where a layer of GRUs RNN was added to the configuration of experiment 1, yielded an increase in performance for all inputs except for the cough.

IV. DISCUSSION

Voice-based screening for COVID19 necessitates a reliable classification of SARS-CoV-2 positives from negatives. The deep learning system in this study aimed to address the challenges of uncertain and presumably subtle vocal attributes of early COVID19, the relatively long sequences in voice recordings and the lack of large labelled datasets. Three core components in the system were an attention-based transformer, GRU-based expert classifiers employing aggressive regularization, and an ensemble stacking. An improved performance was indicated when transformer embeddings were used by the expert classifiers (experiment 3) when compared to the baseline classification of the MEL-spectrum features (experiment 0). Higher F1 values were demonstrated in experiment 3 for all vocal input types in the COVID19 dataset except for the cough inputs. This exception was not surprising, since the transformer was self-trained on speech data and was expected to improve performance for non-speech inputs. Table IV demonstrated that a gradual addition of components to the expert classifiers - RELU layer, GRU and transformer- yielded a gradual improvement in F1 for the three speech inputs. This improvement may imply that the deeper network was able to discover subtle patterns in SARS-CoV-2 positive subjects' voices.

The ensemble stacking in stage 3 generated a meta-model from the expert sub-models that were trained on sub-sets of the data. Table III and Figures 4 and 5 convey that this ensemble stacking resolved a trade-off in the performance of the sub-models. The highest recall performance of 0.8 in two of the sub-models were coupled with high probabilities of false alarm (PFA): 0.45 and 0.56. The meta-model recall of 0.78 was coupled with a lower PFA: 0.41 for the test set and 0.30 for the LOSO cross-validation. A slightly lower recall of 0.76 further decreased PFA to 0.23 (Figure 5). The better performance for the LOSO cross-validation set compared to the test set may indicate that the stacking was more reliable when the SVM had a larger data-set to train on. Furthermore, the LOSO cross validation may have averaged a potential distribution mismatch between the test and train sets. Table III indicated that when using precision instead of PFA as a measure for the false positives, the performance for the test set was better than the performance in the LOSO cross-validation: 0.79 and 0.71 respectively. This trend was also reflected in a higher F1 for the test set: 0.79 compared to 0.74 in the LOSO cross-validation. The performance measures for the test set may be less reliable, however, due to its smaller size.

A small set of vocal-input types were used in this study. The sub-models' performance (Table II) implied that the /z/ phoneme yielded better performance compared to the vocal inputs /ah/ and counting. The /z/ was indicated as relevant in laryngeal disorders discrimination [25]. The counting task was hypothesized to induce and reflect vocal fatigue [27]. The /ah/ utterance was studied in respiratory disorders voice analysis and was associated with sore throat [34]. All three symptoms have been associated with COVID19 [4], [7]. The manifestation of COVID19 symptoms in the vocalization of the phonemes /z/ and /ah/ and of the counting utterance may be explored in clinical voice studies. The acted cough input, on the other hand, demonstrated worse performance compared to all other inputs. This input type may either be less indicative for SARS-Cov-2 detection compared to the other vocal inputs, or may necessitate different classifiers which should be trained or pre-trained on cough-related data.

The expert classifiers in the present study provided F1 values between 0.67 to 0.70 for speech utterances without a transformer pre-training, and between 0.74 and 0.8 when a transformer was used. These F1 values were lower than the ones reported in respiratory disease single speech-input analyses [1]–[3] and may reflect the challenge involved in the detection of COVID19 from voice. Mel-scaled frequency coefficients from free speech recordings were recently used to classify COVID19 symptoms severity [7]. F1 values of 0.65 and 0.66 were obtained, using two different feature sets, in a classification of three-stage symptoms severity. The study differs in its data, paradigm and goals from the ones reported here. It may, however, imply an estimate of the low performance of COVID19 symptoms classification.

The classification in our study used RT-PCR labels as ground truth and therefore reflects a presence of SARS-CoV-2. Information on the subjects' symptoms was not included in the dataset. The correlation between the presence of SARS-CoV-2 detection and COVID19 symptoms in subjects' voice was not investigated previously. The results in our study may only imply a viability for a correlation. A dataset including reliable reports or sensor-based acquisition of symptoms may assess and quantify this correlation.

The COVID19 dataset used in the study was acquired through self-recording using a smartphone application. The preliminary results imply a feasibility for the use of this globally accessible data collection for Sars-COV-2 detection.

Acknowledgment

The authors thank Eli Friedman, Shmuel Makmal, Ido Meister, Shay Perlmutter and Elron Goldemberg from the DDRD who led the data collection and the multi-disciplinary, multi-center, national research project. The authors convey their admiration and gratitude to all doctors, nurses and interns of the Rabin and Sheba medical centers in Israel, who made this unique data collection possible while valiantly fighting the virus. Finally, the authors thank all the participants who recorded themselves and wish them full convalescence and health.

References

[1] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS One*, vol. 12, no. 5, 2017, Art. no. e0177926.

[2] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for COVID-19," 2020, *arXiv:2005.08579*.

[3] C. Bales, C. John, H. Farooq, U. Masood, M. Nabeel, and A. Imran, "Can machine learning be used to recognize and diagnose coughs?" 2020, *arXiv:2004.01495*.

[4] J. R. Lechien *et al.*, "Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus," *J. Intern. Med.*, vol. 288, no. 3, pp. 335–344, Sep. 2020.

[5] X. Zhao *et al.*, "Incidence, clinical characteristics and prognostic factor of patients with COVID-19: A systematic review and meta-analysis," MedRxiv, 2020.

[6] L. Fu, *et al.*, "Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis," *J. Infect.*, vol. 80, pp. 656–665, 2020.

[7] J. Han *et al.*, "An early study on intelligent analysis of speech under covid-19: Severity, sleep quality, fatigue, and anxiety," 2020, *arXiv:2005.00096*.

[8] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, pp. 1735–1780, 1997.

[9] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[10] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," 2019, *arXiv:1904.02874*.

[11] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," *Int. J. Comput. Vis.*, vol. 126, no. 1, pp. 86–110, 2018.

[12] A. Vaswan *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[13] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6419–6423.

[14] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec: Self-supervised learning of discrete speech representations,” presented at the *Int. Conf. Learn. Representations (ICLR)*, 2020.

[15] M. Jiang, J. Wu, X. Shi, and M. Zhang, “Transformer based memory network for sentiment analysis of web comments,” *IEEE Access*, vol. 7, pp. 179942–179953, 2019.

[16] Q. Wang *et al.*, “Learning deep transformer models for machine translation,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1810–1822.

[17] M. Ravanelli *et al.*, “Multi-task self-supervised learning for Robust Speech Recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 6989–6993.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[19] L. Rokach, “Ensemble-based classifiers,” *Artif. Intell. Rev.*, vol. 33, pp. 1–39, 2010.

[20] S. Masoudnia and R. Ebrahimpour, “Mixture of experts: A literature survey,” *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 275–293, Aug. 2014.

[21] Y.-B. Kim, K. Stratos, and D. Kim, “Domain attention with an ensemble of experts,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2017, pp. 643–653.

[22] H. Varian, “Bootstrap tutorial,” *Mathematica J.*, vol. 9, no. 4, pp. 768–775, 2005.

[23] I. Goodfellow, Y. Bengio, and A. Courville, “Regularization for deep learning,” *Deep Learn.*, pp. 216–261, 2016.

[24] M. de Oliveira Rosa, J. C. Pereira, and M. Grellet, “Adaptive estimation of residue signal for voice pathology diagnosis,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan. 2000.

[25] F. C. Eckel and D. R. Boone, “The S/Z ratio as an indicator of laryngeal pathology,” *J. Speech Hearing Disord.*, vol. 46, no. 2, pp. 147–149, May 1981.

[26] G. Van der Meer, Y. Ferreira, and J. W. Loock, “The S/Z ratio: A simple and reliable clinical method of evaluating laryngeal function in patients after intubation,” *J. Crit. Care*, vol. 25, no. 3, pp. 489–492, Sep. 2010.

[27] C. Nanjundeswaran, J. VanSwearingen, and K. V. Abbott, “Metabolic mechanisms of vocal fatigue,” *J. Voice*, vol. 31, no. 3, pp. 378.e1–378.e11, May 2017.

[28] B. McFee *et al.*, “Librosa: Audio and music signal analysis in python,” in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *Proc. NIPS 2014 Workshop Deep Learn.*, 2014.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.

[32] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8609–8613.

[33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, 2015.

[34] Q. Ou, Y. Lu, Q. Huang, and X. Cheng, “Clinical analysis of 150 cases with the novel influenza A (H1N1) virus infection in Shanghai, China,” *Biosci. Trends*, vol. 3, no. 4, pp. 127–130, Aug. 2009.