COVID-19 DETECTION FROM SPEECH IN NOISY CONDITIONS

Shuo Liu[1], *Adria Mallol-Ragolta*[1,2], *Björn W. Schuller*[1,2,3]

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

ABSTRACT

We explore the integration of audio enhancement into a speech-based COVID-19 detection system in an attempt to make speech captured in noisy environments from everyday life useful for the detection of the virus. For this purpose, two multi-task learning approaches are exploited to jointly optimise a front-end speech enhancement model and a subsequent COVID-19 detection model. In comparison to several baseline methods, such as noisy data augmentation, cold cascade of speech enhancement, and COVID-19 models, our proposed solutions are able to recover a substantial percentage of the performance reduction caused by real-world noises. Our best-performing model, which is trained using the synthetic data of the DiCOVA speech corpus and AudioSet environmental backgrounds, can achieve an average AUC of $76.87\,\%$ on the test data covering a wide range of noise intensities, which is over $10\,\%$ better than a COVID-19 model trained with clean audio.

***Index Terms*—** COVID-19 Detection, Multi-Task Learning, Speech Enhancement, Iterative Optimisation

1. INTRODUCTION

The massive testing of the population has been one of the most effective strategies to control the spread of the *Coronavirus Disease* (COVID-19). Nevertheless, the diagnostic tools employed for such purpose —including *Polymerase Chain Reaction* (PCR) and antigen tests— are expensive, time-consuming, and generate a large amount of waste. To overcome this issue, digital health solutions powered with *Artificial Intelligence* (AI) have the potential to offer remote, large-scale, and cost-effective pre-screening tools.

The symptomatology of COVID-19 presents affectations in the human respiratory system. Thus, it seems reasonable to argue that respiratory sounds can contain salient information to detect the virus. In this regard, previous works in the literature explored cough [1, 2], breath [3, 4], and speech [5, 6] signals for the detection of COVID-19 patients. Hand-crafted [7], spectrogram-based [8], or self-supervised learning-based [9] representations have been extracted from the aforementioned respiratory sounds, and exploited with mono- and multi-type [10, 11] approaches.

Voice-based remote pre-screening tools offer users the possibility to analyse a voice sample whenever and wherever they are. The quality of these recordings, which might contain a wide range of background noises, poses a serious threat to the models performance. Herein, we aim to investigate —to the best of the authors' knowledge for the first time– the performance impact of COVID-19 detection models when analysing speech-based samples contaminated with real-world noises, and provide our countermeasures based on speech enhancement techniques.

The remaining of this paper is laid out as follows. Section 2 highlights some related works in the field, while Section 3 presents the methodology followed. In Section 4, we analyse the results obtained from the experiments conducted, and Section 5 concludes the paper.

2. RELATED WORKS

Speech has been discovered as a promising bio-marker for the detection of COVID-19 using machine learning methods [12]. The research covers those using speech alone or as one of the primary sources for the disease detection [13, 14]. The prominent challenge, *Diagnosis of COVID-19 using Acoustics* (DiCOVA) [15], provides a benchmark for comparing audio-based COVID-19 models. However, current speech-based COVID-19 models were typically developed on clean recordings without taking the models' noise robustness into consideration, thereby limiting their practical applications [9, 16].

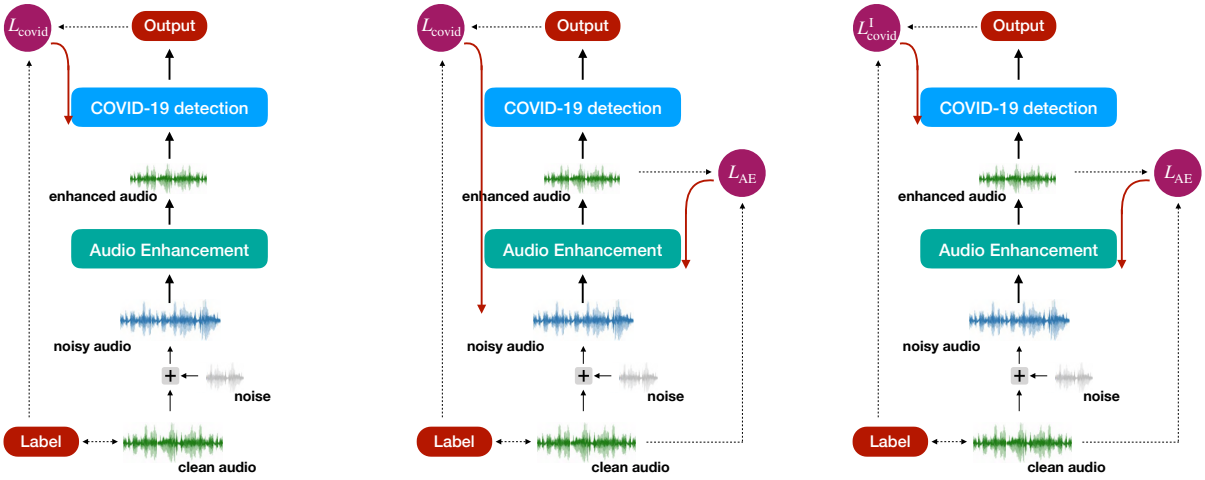
In general, *Audio Enhancement* (AE) can be used as the front-end processing of a computer audio application to improve audio quality in noisy circumstances. The task of *Speech Enhancement* (SE) is typically framed as a supervised learning problem, and its solutions can be broadly categorised as frequency- and time-domain techniques [17, 18, 19]. The frequency-domain solutions either learn a spectral mapping from the *Time-Frequency* (TF) representation of the noisy audio to that of the clean audio, or they estimate a mask that approximates the proposition of the clean component on each TF-bin of the noisy spectrogram [20, 21]. The time-domain SE models, such as waveNet [22] and Wave-U-Net [23], can operate directly on the raw audio waveform while naturally preserving the phase information in the signal during processing.

Inspired by the previous work [24] which investigated the dependence of downstream models on upstream speech enhancement model for reducing the processing artifacts, we choose a U-shaped neural network for speech enhancement [18] that is able to recover audio with high clarity to advance the robustness of a COVID-19 model presented in [11] to real-world noises.

3. METHODOLOGY

In this section, we first present our two multi-task learning paradigms, followed by the definition of several baseline methods for

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), and from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS). It has also been supported by the Bavarian Ministry of Science and Arts through the ForDigitHealth project, funded by the Bavarian Research Association on Healthy Use of Digital Technologies and Media.
Correspondence: shuo.liu@uni-a.de



(a) Cold cascade + data augmentation (b) Multi-task learning (c) Iterative optimisation

Fig. 1: Diagrams showing the methodologies used. The red arrows indicate the back-propagation through the network modules with respect to the losses L of the AE and the COVID-19 detection model.

comparison. Then, we detail the architecture of the neural networks for speech enhancement and COVID-19 detection, respectively.

3.1. Training Paradigms

We explore two joint optimisation methodologies for training an audio enhancement and COVID-19 detection models, i. e., multi-task learning and iterative optimisation, both of which aim to strengthen the mutual promotion between the front-end and the target models.

3.1.1. Conventional Multi-Task Learning

The first approach employs a *Multi-Task Learning* (MTL) framework that combines the losses of the audio enhancement system and the COVID-19 model. The overall loss is mathematically defined as

$$L = L_{\text{AE}} + L_{\text{covid}}. \tag{1}$$

This loss equally weights the losses from both models. Hence, minimising L simultaneously optimises both models.

The difference from a standard MTL problem lies in the alignment and the connection of the two models. Even though the two models are treated as a single entity, the AE loss is derived from an intermediate system layer. Consequently, minimising the AE loss has no influence on the parameters of the COVID-19 model, but the COVID-19 loss back-propagates through the AE model. Therefore, although the AE and the COVID-19 losses function as mutual regularisation terms, they also introduce a bias towards the update of the AE parameters. A similar effect has been observed in previous research on supervised auto-encoders [25].

3.1.2. Iterative Optimisation

Iterative optimisation trains the AE and the COVID-19 models in an iterative manner. The primary motivation behind this method is based on a joint view of the two models. First, the COVID-19 model should always be adapted to the output of the AE model, which may contain residual noise, introduced speech distortions, and artifacts, among others. Second, the performance of the COVID-19 model can be utilised to improve the training process of the AE model, allowing the optimisation to focus on the samples that pose particular challenges to the task of COVID-19 detection. By doing so, we target the optimum performance of the complete neural system, including the front-end audio enhancement and the subsequent COVID-19 detection.

To implement the iterative optimisation, given a batch of samples $x = [x_1, x_2, ..., x_i, ..., x_N]$, we weigh the AE loss by the normalised COVID-19-related loss, such as

$$L^{\text{I}}_{AE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} w_i L_{AE}(x_i, \hat{x}_i), \tag{2}$$

where

$$w_i = L_{covid}(c_i, \hat{c}_i) \tag{3}$$

indicates the importance weights which sum up to 1, and c_i, $\hat{c}_i$ represents true and predicted COVID-19 labels. These weights play the role of sample-level importance to assist in training the AE model, so that the contribution of the conflicting samples – for instance, those corrupted by more intensive noise – is increased.

For training the COVID-19 model, it should process the data coming from the AE system as opposed to the clean signal to prevent the performance gap caused by a cold cascade of the AE and the COVID-19 models.

As long as the AE model is optimised, a more robust COVID-19 model needs to be adapted to the enhanced audio, and a more robust COVID-19 model may further assist the AE model's optimisation by updating the sample difficulties. Therefore, we alternate between the training of the COVID-19 and the AE models; i. e., training the COVID-19 model based on the AE output while freezing the parameters of the AE system, and training the AE system with the indications from the COVID-19 outcomes. The iterative execution of both optimisation steps can gradually approach to an optimum solution.

3.2. Comparison Methods

To assess the efficacy of our proposed joint optimisation methods, we compare them with the methods outlined below over different

levels of noise intensity.

- **Baseline**: COVID-19 model only trained on clean data. Since these models are not optimised for noise robustness, we anticipate a considerable performance decrease when confronted with noisy data.

- **Cold Cascade**: The COVID-19 model exploits a front-end AE component. However, the AE and the COVID-19 models are independently optimised. To do this, the AE model is trained to reach a satisfactory enhancement performance, and then the COVID-19 model is trained with clean data and stacked on top of the AE model.

- **Cold Cascade + *Data Augmentation* (DA)**: The COVID-19 detection model is trained on synthesised noisy data. We intentionally introduce noise into the clean audio recordings with different SNR ratios. The AE model is trained to achieve a satisfactory performance. Due to the models' exposure to noisy data and the incorporation of an AE component, this method should exhibit promising noise robustness.

3.3. Network Architectures

This section describes the U-Net model for audio enhancement and the *Residual Network* (ResNet) for COVID-19 detection. We also detail the adjustments made to connect these two models.

3.3.1. Audio Enhancement U-Net

The audio enhancement U-Net [26] takes the time-frequency representations of an audio signal as input. Only the spectrogram – representing the magnitude of the spectrum – is used, while the decomposed phase spectrum is left unaltered. The network has an auto-encoder architecture with feed-forward layers that stack each encoder layer with its mirrored decoder layer. The encoder analyses the spectrogram of a noisy audio input, and decomposes the audio of interest, for example speech, from the noise components into separate feature maps. The decomposition ability increases with the encoder depth. Then, the decoder recombines necessary feature maps to reconstruct the enhanced audio. Similar to ResNet, skip-connections can facilitate the retrieval of more complete information from the noisy input for the reconstruction of the desired audio.

Given a clean sample x, a spectrogram Y is generated from the contaminated audio y. The U-Net aims to estimate a ratio mask $\mathrm{Mask}(\cdot)$, which is used to filter the original noisy audio and produce the enhanced spectrogram; i. e.,

$$\hat{X} = Y \cdot \mathrm{Mask}(Y). \tag{4}$$

Using the inverse STFT, the enhanced audio $\hat{x}$ can be reconstructed with the phase information of the noisy input. The model parameters are optimised by minimising the *weighted SDR* (wSDR) loss of the original and the estimated clean speech and noise [18]; i. e.,

$$L_{\mathrm{SE}}(x, \hat{x}) = \alpha L_{\mathrm{SDR}}(x, \hat{x}) + (1 - \alpha) L_{\mathrm{SDR}}(n, \hat{n}). \tag{5}$$

In Equation (5),

$$n = y - x \quad \text{and} \quad \hat{n} = y - \hat{x},$$

which represent the actual and the estimated noise signal. Next,

$$L_{\mathrm{SDR}}(x, \hat{x}) = -\frac{< x, \hat{x} >}{||x|| \cdot ||\hat{x}||}, \tag{6}$$

where $< x, \hat{x} >$ indicates the inner product of the actual clean signal and the enhanced output. Finally,

$$\alpha = \frac{||x||^2}{||x||^2 + ||n||^2} \tag{7}$$

is a hyper-parameter used to weight the importance of the audio of interest and the noise during model optimisation.

3.3.2. ResNet-Based COVID-19 Detection

The architecture of the COVID-19 detection model is based on a ResNet-18 model [11], using the pre-trained weights to initialise the network. A dense layer shrinks the embedded representations learnt into a more compact representation, reducing the dimensionality of the output features to 16. The final classification is accomplished using two fully-connected layers with a dropout rate of 0.3. Following the first layer, the output is transformed using a ReLU function, and then fed into the second layer, which implements two output neurons with Softmax to interpret the network outputs as the probability score of the actual sample to correspond to a COVID-19 positive or negative patient. The categorical cross-entropy loss is used to optimise this model.

3.3.3. Systematic Combination

To enable the flexible cascade of the audio enhancement system and a target audio model into a sequence, as well as to train the overall system in an end-to-end fashion, we make a minor but crucial modification to the U-Net specifications for audio enhancement by setting the max-pooling along the time-axis to 1, and leaving the pooling along the frequency-axis unchanged. In this way, the audio enhancement model is able to process audio signals of varying durations. As a consequence, the AE system is now compatible of cascading with the subsequent audio models. Intermediate features are extracted from the enhanced waveform or the AE outcome.

4. EXPERIMENTS

First, we test the robustness of the COVID-19 model against several levels of noise. To do this, we augment the DiCOVA [15] test set with chosen environmental recordings from AudioSet [27]. We then perform speech enhancement using a U-Net independently trained on the created noisy data, with the hypothesis that the enhanced speech would have a higher audio quality, hence enhancing the stability of the COVID-19 detection from speech. Furthermore, by adding environmental noises to the training data of the DiCOVA corpus, we intend to improve the robustness of the COVID-19 model. Finally, we evaluate the performance of our joint optimisation approaches in comparison to these baseline methods.

4.1. Data Description and Processing

The DiCOVA corpus comprises coughing, breathing, and speech recordings collected remotely from individuals with and without COVID-19 [15]. Only the samples containing speech recordings are considered in our investigation. The corpus has its own data partitioning, with 172 confirmed positive individuals out of 965 in the development set, and 71 positive patients out of 471 in the evaluation set.

Table 1: Testing results, AUC (%), using the DiCOVA corpus and selected samples from the AudioSet corpus. DA stands for the method using only data augmentation. MTL represents the proposed multi-task learning solution.

Methods	Inf	25dB	20dB	15dB	10dB	5dB	0dB	Average
Original	81.85	74.16	73.48	69.22	65.69	61.85	56.67	66.84
Cold Cascade	-	70.93	70.70	68.01	65.72	64.99	58.08	66.57
Cold Cascade + DA	-	78.42	76.33	73.65	70.02	68.48	66.74	72.27
MTL	-	**81.73**	80.62	**76.98**	**74.59**	74.45	71.15	76.59
Iterative Optimisation	-	81.35	**81.01**	76.49	74.48	**74.73**	**73.12**	**76.87**

The AudioSet corpus [27] contains more than two million human-labelled 10-second environmental sound clips extracted from YoutTube videos. After excluding all noise recordings labelled as *human sounds* according to the provided AudioSet's ontology, we obtained 16 198 samples for the training set, 636 samples for the development set, and 714 samples for the test set.

To synthesise the noisy samples for training and testing, we mix each speech recording from DiCOVA with an AudioSet sample using an SNR ranging from $0, 5, 10, 15, 20, 25$dB. During training, a random SNR is chosen for synthesising each speech sample in order to maximise the overall generalisation ability of the trained model. At test time, the model performance is assessed in terms of all SNRs considered. As input to the COVID-19 model, the logarithmic values of the spectrogram representation of the speech sample are computed.

4.2. Experimental Settings

From our empirical experience, a batch size of 16 is optimal for training a U-Net for audio enhancement. Thus, the batch size remains constant throughout the experiments presented in this section. The system is optimised using Adam with a learning rate of 0.001. Weight decay is additionally applied to the training for an L2 regularisation effect. During training, as the model input, audio recordings of varying lengths are padded to the length of the longest sample within a batch.

4.3. Evaluation Metrics

As suggested by the DiCOVA challenge, we use the *Area Under the Curve* (AUC) as our performance measure. AUC reveals a classifier's ability to differentiate between two classes, and it summarises the *Receiver Operator Characteristic* (ROC) curve, which illustrates the probability curve of TPR versus FPR at different threshold values. A higher AUC score indicates that a model is more effective at discriminating between the two classes in which the data is distributed.

4.4. Results Analysis

According to [11], the implemented ResNet-18 can get an AUC of $81.85\,\%$ on the clean testing data of DiCOVA (cf. Table 1). This model is, however, susceptible to noise disruption, with even a tiny noise (SNR $= 25$dB), causing an AUC drop of more than $7\,\%$. As the noise rises, the detection performance gradually diminishes until it reaches an AUC of $56.67\,\%$ at the SNR of 0dB.

Applying an independently trained SE model to the front-end of the COVID-19 model cannot improve the average AUC result. In particular, although the front-end enhancement has some favourable effects in circumstances with low SNRs, such as 0 and 5dB, the audio distortions introduced by the SE system can hinder the COVID-19 diagnosis in the cases with high SNRs.

Using the augmented data, i. e., adding noise to the speech data of the DiCOVA training set, the noise robustness of the model can be boosted, yielding an average AUC of $72.27\,\%$, and improving the results across all the SNR conditions. Particularly for the low SNR cases, such as 0dB, the detection performance is improved by more than $10\,\%$.

Both of our two presented joint optimisation approaches, multi-task learning and iterative optimisation, are able to further boost the detection performance, yielding an average AUC of $76.59\,\%$ and $76.87\,\%$, respectively. For high SNR cases, such as 20 and 25dB, both approaches can reach a COVID-19 diagnostic success rate comparable to the performance of the original detection model on the clean test set. The iterative optimisation method surpasses the conventional MTL method in conditions with very low SNR like 0dB, demonstrating its advantage in more noisy environments. Overall, the two solutions jointly optimise the models for audio enhancement and COVID-19 detection, resulting in an AUC performance gains of over $4\,\%$.

5. CONCLUSION

This work explored speech-based COVID-19 detection with a focus on the model's tolerance to noise. To this end, we presented two joint optimisation approaches. Experimental findings support that a task-specific speech enhancement system can efficiently recover the speech signal from noisy recordings to improve COVID-19 identification performance. The particular optimisation of the audio enhancement model towards the COVID-19 task substantially boost the detection performance, producing comparable results to the same COVID-19 model when processing clean audio.

In addition to further research into more efficient neural networks for audio enhancement and COVID-19 detection, more attention should be placed on studying the generalisability of our presented training schema to alternative model architectures. Besides, future studies should incorporate other types of noise – such as voice of unwanted speakers or reverberation – to enable a more robust COVID-19 model for real-world applications. Moreover, our proposed training schema should be deployed and investigated in other computer audition applications, so as to optimise the front-end processing towards the target application.

6. REFERENCES

[1] 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 *Proc. INTERSPEECH*, Brno, Czechia, 2021, pp. 941–945.

[2] R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Transfer learning-based cough representations for automatic detection of COVID-19," in *Proc. INTERSPEECH*, Brno, Czechia, 2021, pp. 436–440.

[3] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proc. KDD*, 2020, pp. 3474–3484.

[4] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," *BMJ Innovations*, vol. 7, no. 2, pp. 356–362, 2021.

[5] S. Deshmukh, M. Al Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting COVID-19 from voice," in *Proc. ICASSP*, Toronto, Canada, 2021, pp. 1055–1059.

[6] 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 *Proc. ICASSP*, Toronto, Canada, 2021, pp. 8328–8332, IEEE.

[7] Z. Mostaani, R Prasad, B. Vlasenko, and M. Magimai-Doss, "Modeling of pre-trained neural network embeddings learned from raw waveform for COVID-19 infection detection," in *Proc. ICASSP*, Singapore, Singapore, 2022, pp. 8482–8486.

[8] A. Mallol-Ragolta, F. B. Pokorny, K. D. Bartl-Pokorny, A. Semertzidou, and B. Schuller, "Triplet loss-based models for COVID-19 detection from vocal sounds," in *Proc. EMBC*, Glasgow, UK, 2022, pp. 998–1001.

[9] X.-Y. Chen, Q.-S. Zhu, J. Zhang, and L.-R. Dai, "Supervised and Self-Supervised Pretraining Based Covid-19 Detection Using Acoustic Breathing/Cough/Speech Signals," in *Proc. ICASSP*, Singapore, Singapore, 2022, pp. 561–565.

[10] 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 *Proc. EMBC*, Guadalajara, Mexico, 2021, pp. 1840–1843.

[11] A. Mallol-Ragolta, H. Cuesta, E. Gomez, and B. Schuller, "Multi-type outer product-based fusion of respiratory sounds for detecting COVID-19," in *Proc. INTERSPEECH*, Incheon, Korea, 2022, pp. 2163–2167.

[12] Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Koike, Xiao Li, Zixing Zhang, Yoshiharu Yamamoto, and Björn W. Schuller, "An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety," in *Proc. INTERSPEECH*, pp. 4946–4950. Shanghai, China, 2020.

[13] Carlo Robotti, Giovanni Costantini, Giovanni Saggio, Valerio Cesarini, Anna Calastri, Eugenia Maiorano, Davide Piloni, Tiziano Perrone, Umberto Sabatini, Virginia Valeria Ferretti, et al., "Machine learning-based voice assessment for the detection of positive and recovered COVID-19 patients," *Journal of Voice*, 2021.

[14] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al., "Sounds of COVID-19: exploring realistic performance of audio-based digital testing," *NPJ digital medicine*, vol. 5, no. 1, pp. 1–9, 2022.

[15] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The second Dicova challenge: Dataset and performance analysis for diagnosis of COVID-19 using acoustics," in *Proc. ICASSP*, Singapore, Singapore, 2022, pp. 556–560.

[16] A. Mallol-Ragolta, S. Liu, and B. Schuller, "COVID-19 Detection Exploiting Self-Supervised Learning Representations of Respiratory Sounds," in *Proc. BHI*, Ioannina, Greece, 2022, 4 pages.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[18] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee, "Phase-aware speech enhancement with deep complex u-net," in *Proc. ICLR*, New Orleans, LA, USA, 2019, 20 pages.

[19] Jiaming Cheng, Ruiyu Liang, Zhenlin Liang, Li Zhao, Chengwei Huang, and Björn Schuller, "A deep adaptation network for speech enhancement: Combining a relativistic discriminator with multi-kernel maximum mean discrepancy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 41–53, 2021.

[20] Shuo Liu, Gil Keren, Emilia Parada-Cabaleiro, and Björn Schuller, "N-HANS: A neural network-based toolkit for in-the-wild audio enhancement," *Multimedia Tools and Applications*, vol. 80, pp. 28365–28389, 2021.

[21] Maximilian Strake, Bruno Defraene, Kristoff Fluyt, Wouter Tirry, and Tim Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6674–6678.

[22] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 3291–3295.

[23] Heitor R. Guimarães, Hitoshi Nagano, and Diego W. Silva, "Monaural speech enhancement through deep wave-U-net," *Expert Systems with Applications*, vol. 158, pp. 113582, 2020.

[24] Deliang Wang Ke Tan, "Improving robustness of deep learning based monaural speech enhancement against processing artifacts," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6914–6918.

[25] L. Le, A. Patterson, and M. White, "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," in *Proc. NeurIPS*, Montreal, Canada, 2018, pp. 107–117.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[27] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 776–780.