Computer Audition for Fighting the SARS-CoV-2 Corona Crisis—Introducing the Multitask Speech Corpus for COVID-19

Kun Qian, *Senior Member, IEEE*, Maximilian Schmitt, Huaiyuan Zheng, Tomoya Koike, *Student Member, IEEE*, Jing Han, *Member, IEEE*, Juan Liu, Wei Ji, Junjun Duan, Meishu Song, Zijiang Yang, *Student Member, IEEE*, Zhao Ren, *Student Member, IEEE*, Shuo Liu, Zixing Zhang, *Member, IEEE*, Yoshiharu Yamamoto, *Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Computer audition (CA) has experienced a fast development in the past decades by leveraging advanced signal processing and machine learning techniques. In particular, for its noninvasive and ubiquitous character by nature, CA-based applications in healthcare have increasingly attracted attention in recent years. During the tough time of the global crisis caused by the coronavirus disease 2019 (COVID-19), scientists and engineers in data science have collaborated to think of novel ways in prevention, diagnosis, treatment, tracking, and management of this global pandemic. On the one hand, we have witnessed the power of 5G, Internet of Things, big data, computer vision, and artificial intelligence in applications of epidemiology modeling, drug and/or vaccine finding and designing, fast CT screening, and quarantine management. On the other hand, relevant studies in exploring the capacity of CA are extremely lacking and underestimated. To this end, we propose a novel multitask speech corpus for COVID-19 research usage. We collected 51 confirmed COVID-19 patients' in-the-wild speech data in Wuhan city, China. We define three main tasks in this corpus, i.e., three-category classification tasks for evaluating the physical and/or mental status of patients, i.e., sleep quality, fatigue, and anxiety. The benchmarks are given by using both classic machine learning methods and state-of-the-art deep learning techniques. We believe this study and corpus cannot only facilitate the ongoing research on using data science to fight against COVID-19, but also the monitoring of contagious diseases for general purpose.**

***Index Terms*—Computer audition, coronavirus disease 2019 (COVID-19), deep learning Internet of Medical Things (IoMT), machine learning,**

This work was supported in part by the Zhejiang Lab's International Talent Fund for Young Professionals under (Project HANAMI), China; in part by the JSPS Postdoctoral Fellowship for Research in Japan under Grant P19081 from the Japan Society for the Promotion of Science (JSPS), Japan; in part by the Grants-in-Aid for Scientific Research under Grant 19F19081 and Grant 20H00569 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan; and in part by the European Union's Horizon 2020 Programme by the Smart Environments for Person-Centered Sustainable Work and Well-Being (SustAGE) under Grant 826506. *(Kun Qian, Maximilian Schmitt, Huaiyuan Zheng, and Tomoya Koike contributed equally to this work.) (Corresponding authors: Kun Qian; Huaiyuan Zheng; Juan Liu; Wei Ji.)*

Kun Qian is with the Group on Audition for Intelligent Medicine, Institute of Engineering Medicine, Beijing Institute of Technology, Beijing 100081, China (e-mail: qian@bit.edu.cn).

Maximilian Schmitt, Meishu Song, Zijiang Yang, Zhao Ren, and Shuo Liu are with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: maximilian.schmitt@informatik.uni-augsburg.de; meishu.song@informatik.uni-augsburg.de; zijiang.yang@informatik.uni-augsburg.de; zhao.ren@informatik.uni-augsburg.de; shuo.liu@informatik.uni-augsburg.de).

Tomoya Koike and Yoshiharu Yamamoto are with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: qian@p.u-tokyo.ac.jp; tommy@p.u-tokyo.ac.jp; yamamoto@p.u-tokyo.ac.jp).

Huaiyuan Zheng is with the Department of Hand Surgery, Wuhan Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zhenghuaiyuan@126.com.

Jing Han is with the Mobile Systems Group, University of Cambridge, Cambridge CB2 1TN, U.K. (e-mail: jh2298@cam.ac.uk).

Juan Liu and Junjun Duan are with the Department of Plastic Surgery, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: liujuan_1018@126.com; drd_surgery107@163.com).

Wei Ji is with the Department of Plastic Surgery, Wuhan Third Hospital and Tongren Hospital of Wuhan University, Wuhan 430072, China (e-mail: ; jiwei1230@foxmail.com).

Zixing Zhang is with GLAM—the Group on Language, Audio, and Music, Imperial College London, London SW7 2BU, U.K. (e-mail: zixing.zhang@imperial.ac.uk).

Björn W. Schuller is with GLAM—the Group on Language, Audio, and Music, Imperial College London, London SW7 2BU, U.K., and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: schuller@ieee.org).

I. Introduction

At THE time of writing this article, coronavirus disease 2019 (COVID-19) is affecting more than 200 countries and regions, with more than 37 million confirmed cases and more than 1 million deaths globally [1]. To combat this unprecedented crisis caused by the virus now officially named as SARS-CoV-2 by the World Health Organization (WHO), scientists across different fields are working together to make efforts in epidemiology prediction, clinical diagnosis and treatment, drug and/or vaccine discovery, social distancing management and monitoring, and further countermeasures. In particular, artificial intelligence (AI) and related signal processing (SP) and machine learning (ML) techniques have shown promising power and potential in the past several months [2], [3]. Moreover, the fast developing and still ongoing changing deep learning (DL) technologies [4] can generate more opportunities when getting more and more available data in the broader COVID-19 research community. Computer vision (CV) and its related techniques are mostly used in current state of the art due to its contribution for a fast and accurate assistive check of the chest CT screening. Li *et al.* proposed a DL model called COVID-19 detection neural network (COVNet), which can achieve a sensitivity of 90.0 % and a specificity of 96.0 % for detecting COVID-19 [5]. The database they collected consisted of 4356 chest CT exams from 3322 patients. Furthermore, more available DL models, e.g., COVID-CAPS [6], COVID-Net [7], and COVID-ResNet [8], were provided to help clinical fast diagnosis and management. These DL models all had achieved encouraging results (with an accuracy more than 90 %) by employing different architectures of the convolutional neural networks (CNNs) [9]. In addition, we can see other contributions toward fighting against COVID-19 by leveraging the power of AI. Ge *et al.* [10] investigated the ML and statistical methods for a data-driven paradigm in the drug discovery for COVID-19. Ong *et al.* [11] found the capacity of ML tools to predict COVID-19 vaccine candidates. Al-qaness *et al.* [12] studied ML models in epidemiology related to COVID-19, i.e., building an ML model to forecast the confirmed cases of the upcoming ten days. Yan *et al.* employed a multi-tree XGBoost algorithm to predict the disease's mortality from a database of blood samples ($n = 485$) collected from COVID-19 patients [13]. They found that lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP) are the selected three biomarkers that can predict the mortality of individual patients more than ten days in advance with more than 90.0 % accuracy. Additionally, with the popularity of advanced technologies in 5G, Internet of Things (IoT), and smart phones, AI-enabled methods can be applied to more public services, such as quarantine management, early diagnosis, and prevention of spread [14]–[16]. Recently, Shuja *et al.* [17] published a comprehensive survey on open access databases for facilitating data-driven methods for COVID-19 research. They give an excellent summary of the existing data modalities and ML/DL models and indicate the challenges in this field.

Nevertheless, the works on exploring computer audition (CA) to fight the COVID-19 spread are largely lacking and underestimated even though its noninvasive and ubiquitous character by nature should indicate a promising potential. To this end, we first gave a perspective in detail about the opportunities and challenges of CA for the COVID-19 research in [18]. Moreover, we propose in this study using real-world data to validate the ideas and show the capacity of CA that makes it ready for joining this battle between humans and virus. The main contributions of this work can be summarized as follows. First, we propose a novel speech corpus, which is named the multitask speech corpus for COVID-19 (MSC-COVID-19). To the best of our knowledge, MSC-COVID-19 is the first multitask database and related investigation on using CA for diagnosis and management of COVID-19 suffers, not only for their physical healthcare but also the mental status. The proposed MSC-COVID-19 can facilitate the emotion-aware Internet of Medical Things (IoMT) for mental state assessment during the pandemic. Second, we conduct a series of benchmark experiments using both classic ML methods and state-of-the-art DL models. The baseline results are intended to be helpful and beneficial for a broad scientific community of combating contagious diseases by leveraging the power of AI and CA. Third, as one of the ongoing advanced research projects focused on data science for COVID-19, it can contribute to other fields in designing methodologies, paradigms, and database establishment and sharing.

The remainder of this article will be organized as follows. First, the background and related work will be introduced in Section II. Then, we describe the details of the database, benchmark methods, and toolkits in Section III. The experimental results will be given in Section IV and followed by a discussion in Section V. Finally, we conclude this work in Section VI.

II. Background and Motivation

CA is defined as an interdisciplinary subject, which involves advanced SP and ML technologies to sense, perceive, process, and synthesize acoustic data for computers [19]. The past decades have witnessed the fast development of CA and its successful applications in the healthcare domain, e.g., heart sound recognition [20] and snore sound classification [21]. As indicated by Schuller *et al.* [18], potential CA-based applications for fighting the ongoing COVID-19 global spread can be summarized by two main directions, i.e., speech and sound analysis.

For speech analysis, it can be highly related to the field of *computational paralinguistics* [22] and the relevant well-documented competitive challenges, e.g., as in the INTERSPEECH computational paralinguistics challenge (ComParE) [23]. Based on the clinical characteristics of the COVID-19 patients [24], one finds fever, dry cough, fatigue, headache, myalgia/arthralgia, and shortness of breath as typical symptoms. Thus, the first thing that comes into one's mind might be the detection of speech under a cold [25]. In the ongoing ComParE 2020 challenge, the continuous assessment of breathing patterns is proposed [26]. Moreover, automatically recognizing speech under a pain symptom [27], [28] could be useful for an early warning. It is also found that COVID-19 patients should have a lack of appetite [29], which can be detected via the eating behavior analysis while speaking [30]. Sleepiness assessment can be implemented in both a binary classification task [31] and a regression estimation task (with Karolinska sleepiness scale) [32]. Considering the high mortality risk among the elderly group (a slightly higher mortality rate in male individuals) [24], age and gender information could be of interest to be identified by speech [33], [34]. Children are not within the high risk group, whereas the relevant long-term effects are still unknown and cannot be overlooked [35]. In particular, infant sounds could be the only acoustical factor for analyzing and understanding their status and behavior [32], [36]. Besides, some comorbidities may lead to high risks of mortality by COVID-19 [24], which can be evaluated by speech analysis if the individuals are suffering from head-and-neck cancer [37], asthma [38], or smoking habits [39]. Apart from the aforementioned individual aspects, social effects by COVID-19 can trigger another issue, e.g., the monitoring, management, and evaluation of the social distancing and quarantine. The social isolation of elderly may generate a serious public mental health issue, which is discussed as an emotion recognition task included in this year's ComParE challenge [26]. Speaker identification and counting could be used for monitoring the social distancing, which can be implemented easily via smartphones [40]. Deception and sincerity [41] can be targeted when a person was sent to quarantine. The detection of speech with or without a mask [26] can also contribute to an efficient social prevention of the COVID-19 spread.

For sound analysis, the sound generated by the human body can be the first thing taken into account. Automatic recognition of coughs [42]–[44] can be used as important early screening marker implemented in smart phone audio applications. Furthermore, CA can be used to analyze and recognize the respiratory sounds and lung sounds of patients with pneumonia [45], which could even be easily observed by the prevalent devices, e.g., smartphones [46]. The snore sound analysis [47], [48], which aims to find the pathological changes in the upper airway, may also facilitate the relevant evaluation of sleep of the COVID-19 patients. The association of the cardiac injury with mortality was found in COVID-19 patients [20], [49], which makes the heart sound recognition task useful in an early monitoring process. Among with others, the sound and audio analysis technologies, such as 3-D audio localization [50] and hearing local proximity [51], can be used for monitoring the social distancing and providing warnings.

A direct inspiration of using CA for the COVID-19 research is to evaluate whether we could develop a diagnosis method less expensive and time consuming than the presently common polymerase chain reaction (PCR) and/or CT chest tests. Imran *et al.* [52] proposed an app to build an AI-enabled preliminary diagnosis method for COVID-19 via cough sounds. They indicate a very promising result with an accuracy above 90 % in an overall recognition of coughs by COVID-19, pertussis, bronchitis, and healthy subjects. These results are quite promising and encouraging, whereas some limitations and constraints still need to be addressed as suggested in [52]. We think that guaranteeing an accurate diagnosis based on collected cough data from COVID-19 patients and finding the distinguishing characteristics between COVID-19 coughs and other coughs are the two most difficult factors that have to be addressed. Besides, a CA-based diagnosis method may not become a gold standard in clinical practice due to PCR and/or CT chest tests being widely used and regarded as a convincing diagnosis method. However, CA-based methods can facilitate a noninvasive, convenient, and cheap real-time monitoring system for both confirmed COVID-19 patients and such individuals who are forced into a quarantine (e.g., 14 days at home/hotel). Not only the physical symptoms (e.g., fever, pain, and fatigue) but also their mental status (e.g., anxiety) are essential for COVID-19-related management in real practice. In particular, for the elderly who are living alone, one may need a 24×7 healthcare system during this global pandemic time. In our recent feasibility study, the elderlies' behavior information can be used to predict their mental status [53]. Motivated by these achievements and opportunity mentioned previously, we want to make a novel exploration of CA for the analysis of speech to recognize COVID-19 and monitor patient wellbeing, which can be considered as another possible modality to be used in a sophisticated AI-based diagnosis, treatment, and management paradigm. A pilot study was shown in [54], which gave a promising preliminary result. However, that study was only validated by simple ML methods without involving the state-of-the-art works. Also, the severity estimation was derived from the number of days in hospitalization with no medical gold standard. In this work, we first introduce the MSC-COVID-19 in a comprehensive way. The successful experiences and open toolkits used in the aforementioned challenges will then be considered and applied, for the first time, to this database. For differences between the one proposed in this study and the early work [54], we briefly summarize as follows. First, we conducted a rigorous preprocessing stage of the audio recordings. Specifically, we filtered some interferences in the low frequency band of the raw audio recordings, which were found to affect the final learning performance of models in our initial experiments. Second, the data partitioning is different. In [54], the experiments were executed in a leave-one-subject-out (LOSO) cross-validation evaluation whereas a train/dev/test partition was established in this study. The LOSO partitioning can render the final performance higher compared to the proposed study. Nevertheless, we think this study contributes to a more standardized way by considering reproducibility aspects and computational effort reality. Third, we excluded the severity task (adopted in [54]) in this study because the annotation of severity in [54] was based on the days of being hospitalized, which cannot be an objective and convincing metric. Overall, we think this proposed database can be suitable for future study usage and is more suitable than the one in [54] to become a future publicly used COVID-19 research resource.

III. Materials and Methods

In this section, we first give the key information of the established database. Then, we introduce the benchmark methods and toolkits used in this study.

A. MSC-COVID-19

1) Data Collection: All the participants involved were informed that their voice data will be used only for research purposes. Their agreements for this study were recorded as one of the five following original speech phrases. The data were collected in-the-wild (Fig. 1): we asked the participants to speak five sentences (with neutral contextual meaning). At the same time, three self-report questions were answered by the participants regarding their *sleep quality*, *fatigue*, and *anxiety*, with a discrete score representing levels 1 to 3. The COVID-19 patients' data were collected from March 20 to March 26 in 2020. All the patients were confirmed by PCR test and CT chest test. We used smartphones (iPhone 6 with 16-GB storage) to record all the patients' voices via the WeChat App.

Following, we give examples of the recorded sentences for COVID-19 patients.

1) 今天是YYYY年MM月DD日。
2) 我同意使用我的语音进行与肺炎相关的研究。
3) 这是我住院的第D天。
4) 我很想早点康复出院。

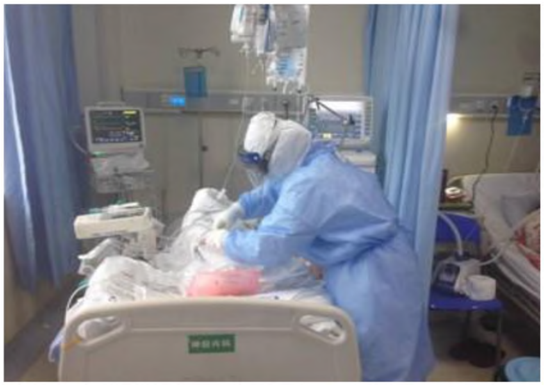


Fig. 1. MSC-COVID-19 database collection environment. The speech data of the confirmed COVID-19 patients are recorded via a smart phone (iPhone 6) by WeChat App.

5) 今天的天气是X。

(translated into English)

1) Today is MM (Month) DD (Day) YYYY (Year).
2) I agree to use my voice for coronavirus related research purposes.
3) Today is the Dth day since I stayed in the hospital.
4) I wish I could rehabilitate and leave the hospital soon.
5) The weather today is X (e.g., sunny).

2) Data Preprocessing: As described in [54], we executed a series of data preprocessing stages before we established the "standard" MSC-COVID-19, which includes *data cleansing*, *voice activity detection*, *speaker diarisation*, and *speech transcription*. First, we excluded recordings of too low quality (e.g., the level of speech is low compared to the background noise). Then, we removed the nonspeech parts from each recording, which results in maintaining only the segments, including voice (e.g., speech, breathing, and coughing) from the recordings. The segments containing solely the target patient and scripted content (e.g., excluding laughing) were kept. Finally, we obtain 260 audio recordings from 51 COVID-19 patients. We understand that the size of the data in the current study is quite limited. But at the same time, this is highly validated data as opposed to the concurrent low-control crowdsourcing efforts rendering this data unique to date. To attenuate the effects of the audio recording equipment, background noise condition, and the level of the recording, all files were first high-pass filtered to eliminate low-frequency background noise (cutoff frequency: 120 Hz, 10th-order Chebyshev filter) and then their waveforms were normalized individually (peak amplitude set to -3 dB).

3) Tasks Definition: We define three tasks for the MSC-COVID-19 benchmark setup. First, we consider three categories of Sleep Quality: Good (labeled as "1"), Normal (labeled as "2"), and Bad (labeled as "3") should be classified from the speech data of COVID-19 patients. Second, the Fatigue Degree should be grouped into: Mild (labeled as "1"), Moderate (labeled as "2"), and Severe (labeled as "3"). Finally, an estimation of the Anxiety Degree should be made as: Mild (labeled as "1"), Moderate (labeled as "2"), and Severe (labeled as "3"). We name these three tasks: **S** (three-class classification), **F** (three-class classification), and **A** (three-class classification) in the following description.

TABLE I
NUMBER [#] OF INSTANCES IN THE DATA PARTITIONS OF MSC-COVID-19. (a) SLEEP QUALITY ESTIMATION TASK. (b) FATIGUE ESTIMATION TASK. (c) ANXIETY ESTIMATION TASK

(a)

	Train	Dev	Test	Σ
Good	41	27	32	100
Normal	31	10	5	46
Bad	74	19	21	114
Total	146	56	58	260

(b)

	Train	Dev	Test	Σ
Mild	22	10	22	54
Moderate	83	22	26	131
Severe	41	24	10	75
Total	146	56	58	260

(c)

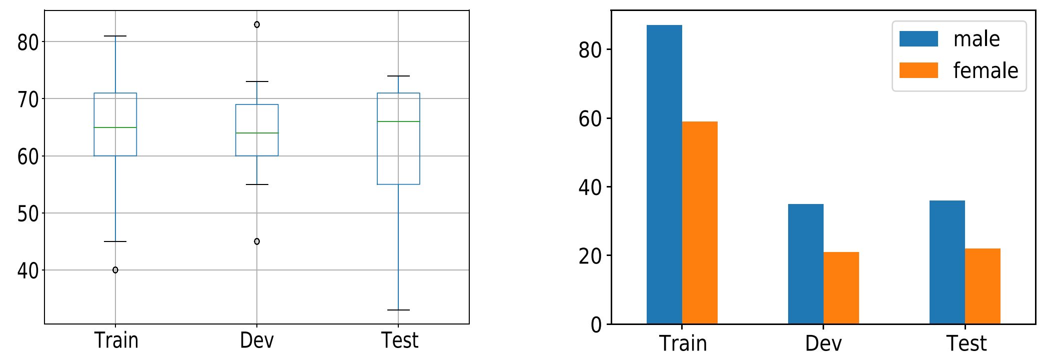
	Train	Dev	Test	Σ
Mild	15	10	16	41
Moderate	99	30	21	150
Severe	32	16	21	69
Total	146	56	58	260

4) Data Partitioning: Considering the gender, age, and annotation distribution (see Fig. 2), we split the overall data into train(ing), dev(elopment), and test sets (Table I). All the ML/DL models' hyperparameters are optimized on the dev set and applied for training the final model on a fusion of the train and dev sets, evaluated on the test set.

B. Benchmark Methods and Toolkits

1) Large-Scale Acoustic Features: In the paradigm of classic ML, features representing acoustic properties are essential for further model building. These features, e.g., Mel-frequency cepstral coefficients (MFCCs), are human hand-crafted needing specific domain knowledge. We use the standard large-scale ComParE [55] feature set in this study extracted by our open-source toolkit openSMILE [56], [57], for its popularity as a standard feature extractor in our previous body of sound analysis tasks, e.g., snore sound [58] and heart sound [20]. The ComParE feature set contains 6373 static features resulting from calculating the statistical *functionals* over low-level descriptors (LLDs) extracted from frames (60-ms size with 10-ms hop size) of the audio files. As a kind of *suprasegmental* features [55], functionals can represent higher statistical information from a given chunk of the signal, and makes the feature set independent of the audio length (see Fig. 3), which is needed for a static classifier, e.g., support vector machine (SVM) [59]. The details of LLDs and the corresponding functionals can be seen in Tables II and III, respectively.

2) Bag-of-Audio-Words Approach: Different from the aforementioned functionals, the Bag-of-Audio-Words (BoAW) approach can extract higher representations from the whole training set per subject rather than only one instance. The term BoAW was derived from the Bag-of-Words (BoW) approach [60], which was successfully



(a) (b)

Fig. 2. Age and gender distribution of MSC-COVID-19. There is no considerable difference between train, dev, and test sets. (a) Distribution of Age. (b) Distribution of Gender.

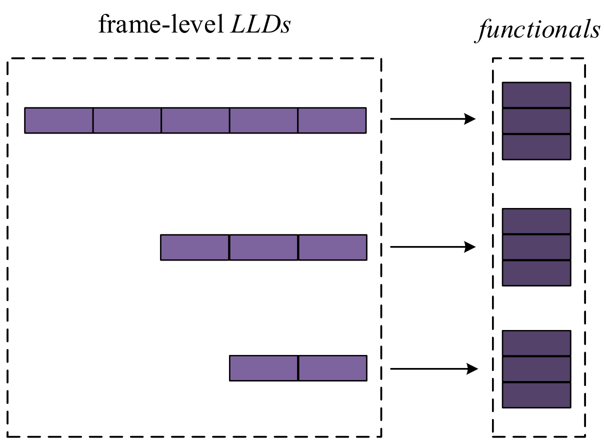


Fig. 3. Scheme of the statistical functionals approach. The frame-level LLDs (e.g., MFCCs) are first extracted from the speech signal. Then, a series of statistical functionals (e.g., max., min., mean, etc.) can be calculated from these LLDs, as scalars independent of the length of the instances.

TABLE II
LLDs FOR COMPARE FEATURE SET. RASTA: RELATIVE SPECTRAL TRANSFORM; HNR: HARMONICS TO NOISE RATIO; RMSE: ROOT MEAN-SQUARE ENERGY; AND SHS: SUBHARMONIC SUMMATION. DETAILS CAN BE FOUND IN [55]

55 Spectral LLDs	**Group**
MFCCs 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filtered auditory spectral bands 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral roll-off point 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 Voicing related LLDs	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Probability of voicing	Voice Quality
log HNR, jitter (local and δ), shimmer (local)	Voice Quality
4 Energy related LLDs	**Group**
RMSE, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic

applied in the domain of *natural language processing* [61] and *computer vision* [62], [63]. Fig. 4 shows the scheme of the chosen BoAW approach. First, a codebook is generated

TABLE III
FUNCTIONALS APPLIED TO LLDs IN THE COMPARE FEATURE SET. NOTE THAT SOME FUNCTIONALS OF THIS TABLE MAY OR MAY NOT BE USED TO ALL OF THE LLDs LISTED IN TABLE II. DETAILS CAN BE FOUND IN [55]

Functionals
Temporal centroid
Peak mean value and distance to arithmetic mean
Mean and standard deviation of peak to peak distances
Peak and valley range (absolute and relative)
Peak-valley-peak slopes mean and standard deviation
Segment length mean, minimum, maximum, standard deviation
Up-level time 25 %, 50 %, 75 %, 90 %
Rise time, left curvature time
Linear prediction gain and coefficients 1–5
Arithmetic or positive arithmetic mean
Root-quadratic mean, flatness
Standard deviation, skewness, kurtosis, quartiles 1–3
Inter-quartile ranges 1–2, 2–3, 1–3,
99-th and 1-st percentile, range of these
Relative position of maximum and minimum value
Range (difference between maximum and minimum values)
Linear regression slope, offset
Linear regression quadratic error
Quadratic regression coefficients
Quadratic regression quadratic error

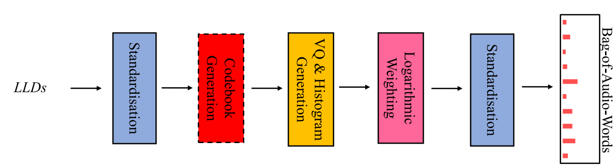


Fig. 4. Processing chain of the BoAW approach. The *term frequency histograms* are regarded as the higher representations extracted from LLDs for further ML/DL models.

from the acoustic LLDs/deltas via a *random sampling* process (the seed is set be a constant to make the study reproducible) following the initialization step of *k-means++ clustering* [64]. Then, each LLD/delta is assigned to the ten audio words from the codebook having the lowest *Euclidean* distance when calculating the histograms. In particular for this study, both BoAW representations from the LLDs and their deltas are concatenated. Finally, a logarithmic term frequency weighting is used to compress the numeric range of the resulting histograms. The LLDs and their corresponding deltas are

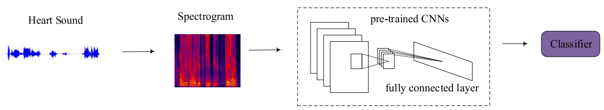


Fig. 5. Scheme of the deep spectrum transfer learning approach. In this paradigm, speech segments are first transformed to spectrograms. Then, a pretrained deep CNN model (e.g., AlexNet) can extract higher representations from these spectrograms. Finally, a classifier (e.g., SVM) can make the predictions based on those higher representations.

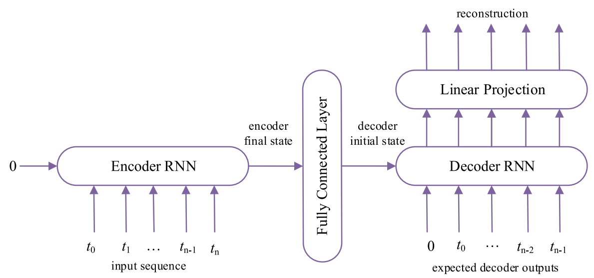


Fig. 6. Scheme of the recurrent autoencoder-based S2SAE approach. In this approach, higher representations are learned in an unsupervised scenario. The process of training the network is to minimize the *root mean-square errors* between the input sequence and the reconstruction. The activations of the fully connected layer are regarded as the high-level representations of the input sequence when the training is complete.

extracted by the OPENSMILE toolkit [57] with the COMPARE feature set as was detailed above. For the BoAW implementation, the OPENXBOW toolkit [65] is used. We investigate 125, 250, 500, 1000, and 2000 for optimizing the codebook size N_c.

3) Transfer Learning: In this transfer learning (TL) [66] paradigm, audio signals are first transformed to spectrograms. Then, the high-level representations of the spectrograms can be extracted from the activations of the fully connected layers of a pretrained deep CNN [9]. Thus, a classifier can perform the classification task by using the extracted high-level representations. Motivated by the previous success of this deep TL method on snore sound [67], heart sound [68], and speech with and w/o mask [69] tasks, we consider investigating it for the MSC-COVID-19 tasks. The speech signals are transformed into Mel-spectrograms (128 Mel frequency bands are computed) using a Hanning window with 32-ms width and 16-ms overlap (Fig. 5). Several kinds of CNN architectures can be employed for high-level representation extraction (the activations of the "avg_pool" layer of the network). Finally, an SVM is used as a classifier to predict the target labels. We investigate ResNet 50 [70], VGG 16 [71], VGG 19 [71], AlexNet [72], and GoogLeNet [73] as pretrained models. The DEEPSPECTRUM [67] toolkit is used for the TL models' implementation.

4) Sequence-to-Sequence Autoencoder Method: In this sequence-to-sequence autoencoder (S2SAE) method (see Fig. 6), the first step is the same as the previously proposed TL method, Mel-scale spectrograms are generated from the speech data. Then, a distinct recurrent S2SAE is trained on each of those sets of spectrograms in an unsupervised scenario, i.e., without any labels. Finally, the learned high-level representations of the a spectrogram are concatenated to form

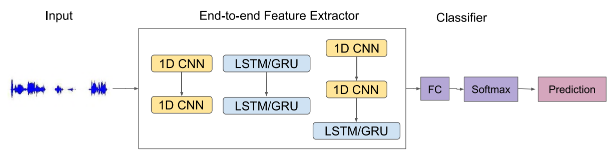


Fig. 7. Scheme of the end-to-end (e2e) learning approach. DL, in essence, is a series of nonlinear transformations of the input. In the paradigm of e2e learning, higher representations can be extracted directly from the raw audio signals. The architecture of the DL models are usually deep CNN and/or RNN models.

the feature vector of the corresponding instance. We use the AUDEEP toolkit [74] to implement the S2SAE method in this study. Furthermore, we evaluate the effects of the background noise, namely, power levels are clipped below certain predefined thresholds (−30, −45, −60, and −75 dB) in the spectrograms.

5) End-to-End Learning: The term e2e can be referred to a holistic paradigm, which connects the input to the output by learned representations from data [75], [76]. In particular, for audio-based applications, it was found that using a CNN to extract features from a waveform can be similar to a Mel-filterbank that is able to automatically discover the frequency decompositions [76]. For automatic speech recognition (ASR), a deep bidirectional long short-term memory (LSTM)-recurrent neural network (RNN) combined with a connectionist temporal classification (CTC) output layer was introduced in [77]. Motivated by the previous success of e2e learning in analysis of music [76], speech emotion [78], and snore sound [79], we investigate several e2e topologies by using a series of CNNs [9] and/or RNNs [80] to extract higher representations directly from the raw sound audio waveforms (see Fig. 7). We use our recent proposed open source DEEPSELF toolkit [81] for the e2e learning models' implementation. To avoid the *vanishing gradient* problem in RNN training [82], we use LSTM [83] and gated recurrent unit (GRU) [84] cells in the deep RNN models.

C. Evaluation Metrics

1) Unweighted Average Recall: To make a fair comparison with the current benchmark and future studies based on MSC-COVID-19, we use the unweighted average recall (UAR) as the main evaluation metric (e.g., to optimize the models' hyperparameters on the dev set). UAR takes the data imbalance characteristics into account [85], which can avoid an overly optimistic evaluation by using the weighted average recall (WAR), i.e., accuracy. Its value is defined as

$$\mathrm{UAR} = \frac{\sum_{i=1}^{N_{\text{class}}} \mathrm{Recall}_i}{N_{\text{class}}} \tag{1}$$

where Recall_i and N_{class} are the Recall of the ith class and the number of classes, respectively. The WAR (accuracy) can be written as

$$\mathrm{WAR} = \sum_{i=1}^{N_{\text{class}}} \lambda_i \mathrm{Recall}_i$$
$$\lambda_i = \frac{N_i}{N} \tag{2}$$

where λ_i is the *weight* for the ith class, N_i is the number of instances labeled as the ith class, and N is the total number of instances.

We also show the confusion matrices of the best models to provide the detailed results. In addition, a significance-level test (one-tailed z-test [86]) is conducted when comparing two algorithms. The results which show a p-value lower than .05 are regarded as significant.

IV. Experimental Results

We will show the experimental results in this section. A brief description of the experimental setup will be given at first.

A. Setup

To make this study reproducible and sustainable, we use exclusively open-source toolkits, including openSMILE [56], [57], openXBOW [65], DeepSpectrum [67], auDeep [74], and deepSELF [81]. All experiments for running these aforementioned toolkits are implemented as Python scripts. For implementing the SVM model, we use the Python *sklearn*[1] toolkit (a *linear kernel* is selected for this study), which is based on the popular LIBSVM toolkit [87]. For training the e2e models, we investigate and compare five topologies, single CNN, single RNN (GRU), and hybrid CNN+RNN (GRU). We also investigated LSTM cells when training the RNNs, whereas their performances yielded to GRU cells in the initial experiments. Therefore, we only use GRU cells in training the deep RNNs, as they tend to be more efficient. The candidates of hyperparameters of single CNNs are 16 and 8 as kernel size, and 16 and 8 as stride size. We also investigate hyperparameters of the RNNs, which are 1 and 2 as the number of RNN layers, and 10 and 50 as the number of hidden nodes. The initialization of all the DL models is generated via randomization (with a constant seed).

All the hyperparameters of the models are tuned and optimized in a grid search strategy on the dev set, and applied to the test set by training the merged data set of train and dev. In the following result part, the dev results are only shown with the optimal ones while the test results are the ones achieved by the optimized model.

TABLE IV

Results for the Benchmarks of the MSC-COVID-19. C: Complexity Parameter of the SVM. N_c: Codebook Size of BoAW Splitting the Input Into Two Codebooks (ComParE-LLDs/ComParE-LLD-Deltas), With Ten Assignments Per Frame, and Optimized Complexity Parameter of the SVM. X: Power Levels That are Clipped Below Four Given Thresholds. N_{e2e}: Number of Layers in the LSTM/GRU/CNN Models for e2e Learning. **UAR**: Unweighted Average Recall. S: Sleep Quality Estimation (Chance Level: 33.3 % of UAR); F: Fatigue Estimation (Chance Level: 33.3 % of UAR); A: Anxiety Estimation (Chance Level: 33.3 % of UAR). The Best Results on the dev and Test Sets are Highlighted in Bold Font. The Best Results on the Test Set are Also Marked With a Gray Background

UAR [%]	**S**		**F**		**A**	
	Dev	Test	Dev	Test	Dev	Test
C	**openSMILE: ComParE func. + SVM**					
10^{-5}	21.7	58.4	47.9	36.4	44.3	56.2
10^{-4}	19.7	62.2	**48.0**	**41.0**	**45.3**	**55.3**
10^{-3}	23.7	49.6	38.6	37.5	44.0	49.5
10^{-2}	**30.2**	**44.0**	38.7	31.4	41.8	38.0
10^{-1}	29.0	48.0	38.7	29.9	41.8	33.2
1	29.0	48.0	38.7	29.7	41.8	33.2
N_c	**openXBOW: ComParE BoAW + SVM**					
125	**36.0**	**33.3**	31.6	31.1	65.4	46.8
250	35.6	34.4	**39.7**	**34.5**	59.7	44.8
500	28.5	36.5	38.0	36.5	**66.7**	**41.1**
1 000	32.3	40.9	31.5	34.2	56.7	45.2
2 000	29.0	35.9	36.6	43.0	56.7	42.2
Network	**DeepSpectrum + SVM**					
AlexNet	**45.7**	**24.7**	34.6	37.1	45.3	39.1
GoogLeNet	37.6	26.8	42.2	30.7	**54.7**	**33.0**
ResNet 50	33.8	45.7	35.0	37.3	46.4	42.7
VGG 16	38.7	27.2	41.2	22.2	38.9	34.6
VGG 19	43.2	40.2	**46.2**	**34.7**	36.5	40.2
X	**auDeep: RNN + SVM**					
$-30\,dB$	38.4	35.5	33.0	32.6	**60.8**	**32.1**
$-45\,dB$	35.4	33.9	34.3	33.6	43.8	30.8
$-60\,dB$	**38.8**	**30.2**	40.4	29.7	58.2	39.2
$-75\,dB$	33.9	41.9	38.7	41.9	39.4	38.4
fused	35.4	30.2	**43.4**	**40.6**	53.9	40.4
Topology	**deepSELF: e2e, N_{e2e}=2**					
CNN	39.2	34.9	**41.5**	**33.3**	35.8	25.4
RNN	49.4	52.6	40.8	38.0	**47.9**	**43.3**
CNN+RNN	**52.0**	**35.1**	40.4	25.4	44.7	27.9
n	**Fusion of n-Best**					
3	–	43.3	–	42.8	–	49.0
4	–	**44.3**	–	42.1	–	**53.7**
5	–	36.0	–	**44.4**	–	43.8

B. Results

The experimental results (UARs) are shown in Table IV and the confusion matrices of the best models are illustrated in Table V. In summary, the best models can reach a UAR of 44.3 %, 44.4 % and 55.3 % for the **S** Task, **F** Task, and **A** Task, respectively. Among these results, one best result is achieved by a single model (**A** Task) while the other two best results are reached by a late fusion (majority vote) strategy of multiple models (**S** Task and **F** Task). We need to note that the current results have shown promising potential for future emotion-aware IoMT applications by considering the current limited data size and difficult annotation.

For the S Task, the best single model is trained by large-scale acoustic features and an SVM classifier. Unlike the performance for the D Task, TL-based models perform worst when compared with other methods (even lower than chancel level). The S2SAE models are also owning UARs lower than 33.3 %, while e2e models and BoAW models are slightly higher or only reaching this level. When looking at the confusion matrix of the best model [see Table V(a)], "Good" is the easiest category to be recognized while "Normal" is the most difficult one (easily to be incorrectly classified as "Bad").

For the F Task, all the models produce higher UARs than chance level (33.3 %). The classic ML model (by large-scale acoustic features and SVM) and the S2SAE methods occupy the first and the second best single model

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

TABLE V

NORMALIZED CONFUSION MATRICES (IN [%]) OF THE BEST MODELS IN EACH TASK ON THE TEST SET. **S** TASK: LATE FUSION OF OPENSMILE: COMPARE FUNC. + SVM, *C*: .01; DEEPSELF: E2E CNN+RNN; OPENXBOW: COMPARE BOAW +SVM, *C*: .01, N_c: 125; AUDEEP: RNN+SVM, *C*: .1, *X*: −60 DB. **F** TASK: LATE FUSION OF OPENSMILE: COMPARE FUNC. +SVM, *C*: .0001; AUDEEP: RNN+SVM,*C*: .1, *X*: FUSED; DEEPSPECTRUM: VGG 19 +SVM, *C*: .01; OPENXBOW: COMPARE BOAW +SVM, *C*: 1.0, N_c: 250; DEEPSELF: E2E CNN, CHANNEL: [3, 6], KERNEL SIZE: [16, 8], STRIDE SIZE: [16, 8], LEARNING RATE: .0001. A TASK: OPENSMILE: COMPARE FUNC. +SVM, *C*: .0001. (a) **S** TASK (UAR = **44.3 %**, CHANCE LEVEL: 33.3 %). (b) **F** TASK (UAR = **44.4 %**, CHANCE LEVEL: 33.3 %). (c) **A** TASK (UAR = **55.3 %**, CHANCE LEVEL: 33.3 %)

(a)

Pred ->	*Good*	*Normal*	*Bad*
Good	50.0	28.1	21.9
Normal	0.0	40.0	60.0
Bad	28.6	28.6	42.9

(b)

Pred ->	*Mild*	*Moderate*	*Severe*
Mild	40.9	40.9	18.2
Moderate	7.7	42.3	50.0
Severe	10.0	40.0	50.0

(c)

Pred ->	*Mild*	*Moderate*	*Severe*
Mild	56.3	25.0	18.7
Moderate	38.1	52.4	9.5
Severe	9.5	33.3	57.1

positions (41.0 % and 40.6 %), respectively. The best results by the BoAW approach and the TL method are comparable (34.5 % versus 34.7 %) and the e2e model's best result has only reached chance level. For the best model [see Table V(b)], "Severe" has the highest recall while "Mild" yields the lowest recall. However, both the two aforementioned categories have a large proportion of instances that are incorrectly recognized as "Moderate," which is easy to be wrongly grouped into "Severe."

For the A Task, the model trained by large-scale acoustic features and SVM classifier reaches the highest UAR (55.3 %). The e2e model reaches a second best single model position when having a UAR of 43.3 % by only using the deep RNN architecture (with GRU cells). Then, the BoAW-based model is the third best single model showing a UAR of 41.1 % while the S2SAE and TL-based models yield only chance level. When looking at the confusion matrix of the best model [see Table V(c)], we may find that "Mild" and "Severe" both have a proportion of instances to be wrongly predicted as "Moderate."

The late fusion of models cannot generate significantly higher results than the best single models. Only for the S and F Tasks, the fused models can have a slight improvement compared to the best single models.

V. DISCUSSION

We now give a discussion on findings, limitations, and perspectives of this study.

A. First Findings

It is encouraging to see that our proposed CA-based models have a good performance in monitoring the physical and/or mental status of the COVID-19 patients. On the one hand, as we indicated in our preliminary surveys [18] and studies [54], CA-based methods should have a promising capacity in helping diagnosis, precaution, and management of the COVID-19 epidemic. On the other hand, we should not be overoptimistic due to one possible factor that could be leading to such good performances. The MSC-COVID-19 database has a comparably high quality based on a complicated human involved preprocessing step. Nevertheless, in real clinical or daily life practice, it cannot be obtained in such an ideal condition. We should consider more advanced technologies to eliminate the noises, interference, and reverberations.

For all the tasks, the best final results (the baselines) are significantly higher than the corresponding chance level ($p < 0.05$ by one-tailed z-test). For the classic ML models, specifically for large-scale acoustic features trained models (see Table IV), the results are robust for multiple tasks in this study. It can be noted that as observed in this preliminary investigation, human hand-crafted features (with clear definitions and physical meanings) are worth exploring. In addition, limited to the current data size, the DL-based models may have been restrained in their capacities in learning more generalized features.

Management and daily monitoring of the patients' physical and mental status is a crucial task. We are encouraged by the current results (even though not perfect, yet) for using voices to estimate sleep quality, fatigue, and anxiety degrees. In particular, we have seen that even when only using the deep RNN (with GRU cells) architecture and the audio waveform as the input, one can reach a UAR of 43.3 % (as the second best single model) for the A Task. For the S and F Tasks, a late fusion has resulted in a slight improvement, which is worth further studying. For these three computational paralinguistics analysis tasks, the COMPARE feature set shows good robustness due to its design in the context of its original target usage.

B. Limitations and Perspectives

First, the fundamental investigation of the relationship between the acoustic features and the pathological characteristics of COVID-19 is still lacking. Before giving any solid conclusion, we need to collect a larger size of COVID-19 patients' speech data. Additionally, the anthropometric parameters and the ethnics of the patients should be taken into account. We believe that as a global crisis, COVID-19 cannot be beaten by only one single country or one single field of science. In the future, we aim to consider collecting the voice data globally and discover the characteristics of COVID-19 patients' voices internationally.

Second, more advanced SP techniques should be introduced. Similar to our previous findings in snore sound studies [47], [48], wavelet transformation-based features can be superior in multiresolution analysis to the Fourier transformation-based features, which occupy the main part of the COMPARE feature set. Besides, one should consider exploring the learned features by DL models by introducing attention mechanisms [88].

Third, *data scarcity* is a challenging issue for almost all of the health-related AI applications. In future work, we should investigate the ML strategies of unsupervised learning [89], semisupervised learning [90], active learning [91]–[93], and cooperative learning [94], to enrich the COVID-19 speech corpus. We should also consider introducing generative adversarial networks (GANs) to generate more sample instances with a reasonable distribution [95], [96].

Last but not least, to build an explainable AI (XAI) system [97] for CA-based COVID-19 detection and management usage, we need to reach a close collaboration of experts from a multidisciplinary background, including medicine and acoustics.

VI. CONCLUSION

We introduced a novel multitask speech corpus (MSC-COVID-19) for COVID-19 research in this study. To the best of our knowledge, MSC-COVID-19 is the first comprehensive CA-based database that can be used for COVID-19 research purpose. Benchmarks using both classic ML and state-of-the-art DL methods have shown promising preliminary results of using CA for fighting against COVID-19. In particular, we explored the feasibility to evaluate the patients' physical and/or mental status from their voices. We believe that CA-based methods have a great potential to develop noninvasive, cheap, and convenient intelligent systems and/or smart devices to help cope with the crisis caused by contagious diseases. In future work, these proposed multitask CA learning technologies for emotion-aware assessment should be implemented as smartphone apps or embedded in existent ambient audio intelligence connected to the Internet.

ACKNOWLEDGMENT

This database will be released in near future for public research purpose after receiving the approval by Wuhan Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, China. The authors also want to express our deepest sorrow for those who left them due to COVID-19; they are lives, not numbers. They further express our highest gratitude and respect to the clinicians and scientists, and anyone else these days helping to fight against COVID-19, and at the same time help them maintain our daily lives.

REFERENCES

[1] Johns Hopkins University, USA. (2020). *COVID-19 Case Tracker Follow Global Cases and Trends (Updated Daily)*. [Online]. Available: https://coronavirus.jhu.edu/

[2] N. Peiffer-Smadja, R. Maatoug, F.-X. Lescure, E. D'Ortenzio, J. Pineau, and J.-R. King, "Machine learning for COVID-19 needs global collaboration and data-sharing," *Nat. Mach. Intell.*, vol. 2, pp. 293–294, May 2020.

[3] M. Luengo-Oroz *et al.*, "Artificial intelligence cooperation to support the global response to COVID-19," *Nat. Mach. Intell.*, vol. 2, pp. 295–297, May 2020.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[6] P. Afshar *et al.*, "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognit. Lett.*, vol. 138, pp. 638–643, Oct. 2020.

[7] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *Sci. Rep.*, vol. 10, pp. 1–12, Oct. 2020.

[8] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID-19 from radiographs," 2020. [Online]. Available: arXiv:2003.14395

[9] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. NIPS*, Denver, CO, USA, 1989, pp. 396–404.

[10] Y. Ge *et al.* (2020). *A Data-Driven Drug Repositioning Framework Discovered a Potential Therapeutic Agent Targeting COVID-19*. [Online]. Available: https://www.biorxiv.org

[11] E. Ong *et al.*, "COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning," *Front. Immunol.*, vol. 11, pp. 1–13, Jul. 2020.

[12] M. A. A. Al-Qaness, A. A. Ewees, H. Fan, and M. A. E. Aziz, "Optimization method for forecasting confirmed cases of COVID-19 in China," *J. Clin. Med.*, vol. 9, no. 674, pp. 1–15, 2020.

[13] L. Yan *et al.*, "An interpretable mortality prediction model for COVID-19 patients," *Nat. Mach. Intell.*, vol. 2, pp. 283–288, May 2020.

[14] A. S. S. Rao and J. A. Vazquez, "Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine," *Infection Control Hospital Epidemiol.*, vol. 41, no. 7, pp. 826–830, 2020.

[15] S. Latif *et al.*, "Leveraging data science to combat COVID-19: A comprehensive review," *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 85–103, Aug. 2020.

[16] K. Qian *et al.*, "Artificial intelligence Internet of Things for the elderly: From assisted living to health-care monitoring," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 1–11, Jul. 2021.

[17] J. Shuja *et al.*, "COVID-19 open source data sets: A comprehensive survey," *Appl. Intell.*, vol. 51, pp. 1296–1325, Sep. 2021.

[18] B. W. Schuller *et al.*, "COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 Corona crisis," *Front. Digit. Health*, to be published.

[19] K. Qian *et al.*, "Computer audition for healthcare: Opportunities and challenges," *Front. Digit. Health*, vol. 2, no. 5, pp. 1–4, 2020.

[20] F. Dong *et al.*, "Machine listening for heart status monitoring: Introducing and benchmarking HSS—The heart sounds Shenzhen corpus," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2082–2092, Jul. 2020.

[21] K. Qian *et al.*, "Can machine learning assist locating the excitation of snore sound? A review," *IEEE J. Biomed. Health Informat.*, early access, Jul. 29, 2020, doi: 10.1109/JBHI.2020.3012666.

[22] B. W. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.

[23] B. W. Schuller *et al.*, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Comput. Speech Lang.*, vol. 53, pp. 156–180, Jan. 2019.

[24] W. Guan *et al.*, "Comorbidity and its impact on 1590 patients with COVID-19 in China: A nationwide analysis," *Eur. Res. J.*, vol. 55, no. 5, pp. 1–14, 2020.

[25] B. W. Schuller *et al.*, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3442–3446.

[26] B. W. Schuller *et al.*, "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 2042–2046.

[27] Y. Oshrat *et al.*, "Speech prosody as a biosignal for physical pain detection," in *Proc. Speech Prosody*, Boston, MA, USA, 2016, pp. 420–424.

[28] Z. Ren *et al.*, "Evaluation of the pain level from speech: Introducing a novel pain database and benchmarks," in *Proc. ITG Conf. Speech Commun.*, Oldenburg, Germany, 2018, pp. 56–60.

[29] S. Luo, X. Zhang, and H. Xu, "Don't overlook digestive symptoms in patients with 2019 novel coronavirus disease (COVID-19)," *Clin. Gastroenterol. Hepatol.*, vol. 18, no. 7, pp. 1636–1637, 2020.

[30] D. M. Schuller and B. W. Schuller, "The challenge of automatic eating behaviour analysis and tracking," in *Recent Advances in Intelligent Assistive Technologies: Paradigms and Applications* (Intelligent Systems Reference Library), H. N. Costin, B. W. Schuller, and A. M. Florea, Eds. Basel, Switzerland: Springer, 2020, pp. 187–204.

[31] B. W. Schuller *et al.*, "The INTERSPEECH 2011 speaker state challenge," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3201–3204.

[32] B. W. Schuller *et al.*, "The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2378–2382.

[33] B. W. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2794–2797.

[34] F. Weninger, E. Marchi, and B. W. Schuller, "Improving recognition of speaker states and traits by cumulative evidence: Intoxication, sleepiness, age and gender," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 1159–1162.

[35] X. Li *et al.* (2020). *A Mini Review on Current Clinical and Research Findings for Children Suffering From COVID-19*. 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.30.20044545v1

[36] B. W. Schuller *et al.*, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 122–126.

[37] A. Maier *et al.*, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 1–7, Aug. 2009.

[38] I. Mazić, M. Bonković, and B. Džaja, "Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds," *Biomed. Signal Process. Control*, vol. 21, pp. 105–118, Aug. 2015.

[39] H. Satori *et al.*, "Voice comparison between smokers and non-smokers using HMM speech recognition system," *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 771–777, 2017.

[40] C. Xu *et al.*, "Crowd++ unsupervised speaker count with smartphones," in *Proc. UbiComp*, Zurich, Switzerland, 2013, pp. 43–52.

[41] B. W. Schuller *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 2001–2005.

[42] S. Matos *et al.*, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1078–1083, Jun. 2006.

[43] T. Olubanjo and M. Ghovanloo, "Tracheal activity recognition based on acoustic signals," in *Proc. IEEE EMBC*, Chicago, IL, USA, 2014, pp. 1436–1439.

[44] J. Schröder, J. Anemiiller, and S. Goetze, "Classification of human cough signals using spectro-temporal Gabor filterbank features," in *Proc. IEEE ICASSP*, Shanghai, China, 2016, pp. 6455–6459.

[45] R. L. Murphy *et al.*, "Automated lung sound analysis in patients with pneumonia," *Respiratory Care*, vol. 49, no. 12, pp. 1490–1497, 2004.

[46] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *Proc. IEEE IJCNN*, 2015, pp. 1–8.

[47] K. Qian *et al.*, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1731–1741, Aug. 2017.

[48] K. Qian *et al.*, "A bag of wavelet features for snore sound classification," *Ann. Biomed. Eng.*, vol. 47, no. 4, pp. 1000–1011, 2019.

[49] K. Qian *et al.*, "Deep wavelets for heart sound classification," in *Proc. ISPACS*, Taipei, Taiwan, 2019, pp. 1–2.

[50] S. Delikaris-Manias *et al.*, "3D localization of multiple audio sources utilizing 2d doa histograms," in *Proc. IEEE EUSIPCO*, Budapest, Hungary, 2016, pp. 1473–1477.

[51] F. B. Pokorny *et al.*, "Sound and the city: Current perspectives on acoustic geo-sensing in urban environment," *Acta Acustica United Acustica*, vol. 105, no. 5, pp. 766–778, 2019.

[52] A. Imran *et al.*, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informat. Med. Unlocked*, vol. 20, 2020, Art. no. 100378.

[53] K. Qian *et al.*, "Can appliances understand the behaviour of elderly via machine learning? A feasibility study," *IEEE Internet Things J.*, early access, Dec. 15, 2020, doi: 10.1109/JIOT.2020.3045009.

[54] J. Han *et al.*, "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 4946–4950.

[55] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. Cham, Switzerland: Springer Int., 2015.

[56] F. Eyben, M. Wöllmer, and B. W. Schuller, "openSMILE—The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, Florence, Italy, 2010, pp. 1459–1462.

[57] F. Eyben *et al.*, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[58] C. Janott *et al.*, "Snoring classified: The Munich–Passau snore sound corpus," *Comput. Biol. Med.*, vol. 94, pp. 106–118, Mar. 2018.

[59] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[60] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.

[61] F. Weninger, P. Staudt, and B. W. Schuller, "Words that fascinate the listener: Predicting affective ratings of on-line lectures," *Int. J. Distance Educ. Technol.*, vol. 11, no. 2, pp. 110–123, 2013.

[62] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 591–606, Apr. 2009.

[63] J. Wu, W.-C. Tan, and J. M. Rehg, "Efficient and effective visual codebook generation using additive kernels," *J. Mach. Learn. Res.*, vol. 12, pp. 3097–3118, Nov. 2011.

[64] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM SODA*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[65] M. Schmitt and B. W. Schuller, "openXBOW-Introducing the Passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, vol. 18, no. 96, pp. 1–5, 2017.

[66] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[67] S. Amiriparian *et al.*, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.

[68] T. Koike *et al.*, "Audio for audio is better? An investigation on transfer learning models for heart sound classification," in *Proc. EMBC*, 2020, pp. 74–77.

[69] T. Koike *et al.*, "Learning higher representations from pre-trained deep models with data augmentation for the ComParE 2020 challenge mask task," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 1–5.

[70] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–14.

[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[73] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1–9.

[74] M. Freitag *et al.*, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, 2017.

[75] Y. LeCun *et al.*, "Off-road obstacle avoidance through end-to-end learning," in *Proc. NeurIPS*, 2006, pp. 739–746.

[76] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 6964–6968.

[77] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, Beijing, China, 2014, pp. 1764–1772.

[78] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shangai, China, 2016, pp. 5200–5204.

[79] M. Schmitt and B. W. Schuller, "End-to-end audio classification with small datasets–Making it work," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[80] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[81] T. Koike *et al.*, "deepSELF: An open source deep self end-to-end learning framework," 2020. [Online]. Available: arXiv:2005.06993v1

[82] S. Hochreiter *et al.*, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, J. F. Kolen, and S. C. Kremer, Ed. Piscataway, NJ, USA: IEEE Press, 2001, pp. 237–244.

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] J. Chung *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS DLRL Workshop*, 2014, pp. 1–9.

[85] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.

[86] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[87] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011, [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[88] Z. Ren *et al.*, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 113–117.

[89] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.

[90] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[91] B. Settles, "Active learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 6, no. 1, pp. 1–114, 2012.

[92] K. Qian *et al.*, "Active learning for bird sounds classification," *Acta Acustica united with Acustica*, vol. 103, no. 3, pp. 361–364, 2017.

[93] K. Qian *et al.*, "Active learning for bird sound classification via a kernel-based extreme learning machine," *J. Acoust. Soc. America*, vol. 142, no. 4, pp. 1796–1804, 2017.

[94] Z. Zhang *et al.*, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[95] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[96] Z. Zhang *et al.*, "Snore-GANs: Improving automatic snore sound classification with synthesized data," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 300–310, Jan. 2020.

[97] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

Kun Qian (Senior Member, IEEE) received the Doctoral degree for his study on automatic general audio signal classification in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany, in 2018.

Since 2021, he has been appointed as a Full Professor with a title of "Teli Young Fellow" with the Beijing Institute of Technology, Beijing, China. He has a strong collaboration connection to prestigious universities in Germany, U.K., Japan, Singapore, and USA. He has (co)authored more than 70 publications in peer reviewed journals, and conference proceedings having received more than 1,000 citations (H-index 19).

Prof. Qian serves as an Associate Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, *Frontiers in Digital Health*, and *Bio Integration*.

Maximilian Schmitt received the Diploma degree in electrical engineering (Dipl.-Ing.) from RWTH Aachen University, Aachen, Germany, in 2012. He is currently pursuing the Ph.D. degree in computer science with the University of Augsburg, Augsburg, Germany.

He has worked with the Erich-Thienhaus-Institut, University of Music Detmold, Detmold, Germany, and the Computer Science Department, University of Passau, Passau, Germany. He (co)authored more than 50 publications in peer-reviewed journals and conference proceedings, having received more than 800 citations (H-index 14). His research focuses on signal processing, machine learning, and intelligent audio analysis.

Huaiyuan Zheng received the doctoral degree (M.D.) from Wuhan Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013.

He has been working as a Research Fellow for his research on chronic wound healing in Klinikum recuts der Isar, Technische Universität München, Munich, Germany, from 2013 to 2015. He is currently working as a Physician with the Department of Hand Surgery, Wuhan Union Hospital, Tongji Medical College, HUST. As the first or corresponding author, he has more than 15 publications in peer-reviewed journals leading to more than 140 citations. His main research interests include malignant tumor from soft tissue, tissue engineering and regenerative medicine, and medicine-related artificial intelligence. As a front-line doctor, he participated in the battle against the COVID-19 epidemic for more than three months in Wuhan, the capital city of China's Hubei Province.

Tomoya Koike (Student Member, IEEE) received the B.Sc. degree from Kobe University, Kobe, Japan, in 2020. He is currently pursuing the master's degree with the Graduate School of Education, The University of Tokyo, Tokyo, Japan.

He has working experiences with AI-related projects on research and development. He is the main author of the open-source toolkit DEEPSELF. His research interests include machine learning, deep learning, and healthcare applications.

Jing Han (Member, IEEE) received the bachelor's degree in electronic and information engineering from Harbin Engineering University, Harbin, China, in 2010, the master's degree from Nanyang Technological University, Singapore, in 2014, and the Doctoral degree from the University of Augsburg, Augsburg, Germany, in 2020.

She is currently working as a Postdoctoral Researcher with the University of Cambridge, Cambridge, U.K., involved in the ERC project EAR. She has (co)authored more than 30 publications in peer-reviewed journals and conference proceedings, having received more than 580 citations (H-index 16). She reviews regularly for the IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. Her research interests are related to deep learning for affective computing and digital health.

Juan Liu received the Doctoral degree (Ph.D.) for her research on 3-D printing and tissue engineering in the program of medical life science and technology from Technische Universität München, Munich, Germany, in 2018.

She is currently working as a Physician with the Department of Plastic Surgery, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. As the first or corresponding author, she has more than 20 publications in peer-reviewed journals, books, and conference proceedings leading to more than 230 citations. Her main research interests include 3-D printing, tissue engineering and regenerative medicine, and medicine-related artificial intelligence.

Wei Ji received the master's and Doctoral degrees in hand surgery from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2012 and 2015 respectively.

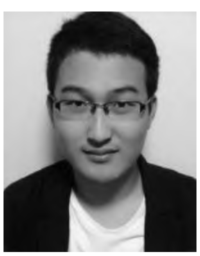
He is currently an Attending Doctor with the Department of Hand Surgery, Wuhan Union Hospital, Tongji Medical College, HUST. As a front-line doctor, he participated in the battle against the COVID-19 epidemic for more than three months in Wuhan, the capital city of China's Hubei Province.

Junjun Duan received the master's degree in plastic surgery from Jianghan University, Wuhan, China, in 2013.

She is currently working as a Physician with the Department of Plastic Surgery, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan. She dedicates to research hypertrophic scar and keloid. Her main research work includes treatments of pathological scar, skin photoaging, Ance vulgaris, and sensitive skin.

Meishu Song received the master's degree in human–computer interaction from the University of York, York, U.K., in 2016. She is currently pursuing the Doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Augsburg, Germany.

Her research interests are emotional intelligence, educational intelligence, and machine learning.

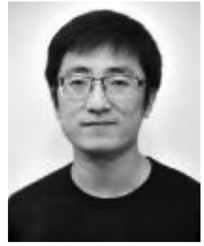


Zijiang Yang (Student Member, IEEE) received the master's degree in information technology from the University of York, York, U.K., in 2016. He is currently pursuing the Doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Augsburg, Germany.

His research focuses on signal processing and analysis, artificial neural networks, and machine learning.

Zhao Ren (Student Member, IEEE) received the master's degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, 2017.

She is currently a Research Assistant and working on her Doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, where she is involved with the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project TAPAS, for emotion analysis based on pathological speech. Her research interests mainly lie in transfer learning, attention mechanisms, and deep learning for the application in healthcare and wellbeing.



Shuo Liu received the B.Sc. degree in network engineering from Nanjing University of Posts and Telecommunication, Nanjing, China, in 2012, and the M.Sc. degree in electrical engineering and information technology from the Technical University of Darmstadt, Darmstadt, Germany, in 2017. He is currently pursuing the Doctoral degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany.

His current research interests include deep learning and machine learning algorithms for speech and audio processing, affective computing, and health-related applications.

Zixing Zhang (Member, IEEE) received the master's degree in physical electronics from Beijing University of Posts and Telecommunications, Beijing, China, 2010, and the Ph.D. degree in engineering from the Machine Intelligence and Signal Processing Group, Technische Universität München, Munich, Germany, 2015.

He was a Research Associate with Imperial College London, London, U.K. He (co)authored more than 90 publications in peer-reviewed journals and conference proceedings (more than 2700 citations, H-index 28). His research interests mainly lie in semisupervised learning, active learning, and deep learning for the application in affective computing.

Yoshiharu Yamamoto (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in education from the University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1990, respectively.

Since 2000, he has been a Professor with the Graduate School of Education, The University of Tokyo, where he is teaching and researching physiological bases of health sciences and education. He is also the President of the Healthcare IoT Consortium, Japan. He (co)authored more than 230 publications in peer-reviewed books, journals, and conference proceedings leading to more than 11 000 citations (H-index 55). His research interests include biomedical signal processing, nonlinear and statistical biodynamics, and health informatics.

Prof. Yamamoto is currently an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and an Editorial Board Member of the Technology and Biomedical Physics and Engineering Express.

Björn W. Schuller (Fellow, IEEE) received the Diploma degree in electrical engineering and information technology, the Doctoral degree in electrical engineering and information technology, and the habilitation and adjunct teaching professorship in the subject area of signal processing and machine intelligence (electrical engineering and information technology) from Technische Universität München, Munich, Germany, in 1999, 2006, and 2012.

He is a tenured Full Professor heading the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, and a Professor with Artificial Intelligence heading GLAM, Department of Computing, Imperial College London, London, U.K. He has (co)authored five books and more than 900 publications in peer-reviewed books, journals, and conference proceedings leading to more than 36 000 citations (H-index 86).

Prof. Schuller is the Field Chief Editor of *Frontiers in Digital Health*, former Editor-in-Chief of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the President-emeritus of the AAAC, the Golden Core Awardee of the IEEE Computer Society, a Fellow of ISCA, and a Senior Member of ACM.