AN OVERVIEW OF THE FIRST ICASSP SPECIAL SESSION ON COMPUTER AUDITION FOR HEALTHCARE

Kun Qian[1], Tanja Schultz[2], and Björn W. Schuller[3,4]*

[1]School of Medical Technology, Beijing Institute of Technology, China
[2]Cognitive Systems Lab, University of Bremen, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[4]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
qian@bit.edu.cn

ABSTRACT

Audio has been increasingly used as a novel digital phenotype that carries important information of the subject's health status. We can find tremendous efforts given to this young and promising field, i. e., computer audition for healthcare (CA4H), whereas the application scenarios have not been fully studied as compared to its counterpart in medical areas, computer vision. To this end, the first special session held at ICASSP 2020 was dedicated to the topic. In this overview paper, we at first summarise the invited high-quality contributions from leading scientists from a multi- disciplinary background. Then, we provide a detailed grouping of the contributions to several scenarios such as body sound analysis (e. g., heart sound), human speech analysis (e. g., stress detection), and artificial hearing technologies (e. g., cochlear implants). In addition to the collected works, we will compare them with other recent studies within the topic. Finally, we conclude the limitations and perspectives of the current stage. It is interesting and encouraging to find that the state-of-the-art machine learning and audio signal processing techniques have been successfully applied in the health domain, e. g., to fight with the global challenges of COVID-19 and ageing population.

Index Terms— Computer Audition, Digital Phenotype, Healthcare, Intelligent Medicine, Overview

1. INTRODUCTION

Computer audition (CA), compared to its counterpart–computer vision (CV), has been underestimated for a long time in the field of digital health. Nevertheless, audio has been increasingly studied as a novel digital phenotype for measuring human beings' health status both physically and psychologically [1]. Benefiting from the ubiquitous equipment (e. g., a microphone/smartphone), audio can facilitate the development of low-cost, convenient, and non-obstructive intelligent medical apparatus that can be used at anytime and anywhere. Besides, passively collecting audio data can overcome the drawback of wearable devices, which might be difficult for elderly individuals to carry such all day [2].

Generally speaking, there are two main directions in the field of CA for healthcare (CA4H): Physical and mental diseases (see Fig. 1). For physical disorders, audio can be used as a kind of physiological signals that carries important information about the subject's body status. Snore sound, as an example, has been demonstrated to be effective to reflect the structure of the upper airway [4]. This finding has been used to diagnose chronic serious sleep disorders, e. g., obstructive sleep disorder (OSA) [5] and/or locating the snore site positions of the subjects [6]. For mental diseases, speech can for example be used as a signal to assess depression level and the suicide risk [7]. On the one hand, there are a plethora of efforts that have been invested into the field of using audio as a novel digital phenotype to measure a subject's physical and mental health status. On the other hand, comparably little attention have been paid by the community of audio and speech processing even though encouraging results were achieved. To this end, we introduced the concept of CA4H at ICASSP 2021 and invited leading individuals in this field to present their most recent state-of-the-art works. To the best of our knowledge, it was the first time to organise a special session at ICASSP on the specific topic of CA applications for healthcare. The main contributions of this overview can be summarised as: We give an overall introduction of the collected contributions at the first ICASSP Special Session on CA4H. This brief can be a good guidance for an audience who shares similar interests and may benefit from a tutorial to start with. Second, readers, particularly those who are not familiar with the relevant field, can enrich their knowledge about the application scenarios. Third, perspectives and insights from the authors shall benefit the field's future work.

The remainder of this paper will be organised as follows:

This work was supported by the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China. *Corresponding author*: Kun Qian.

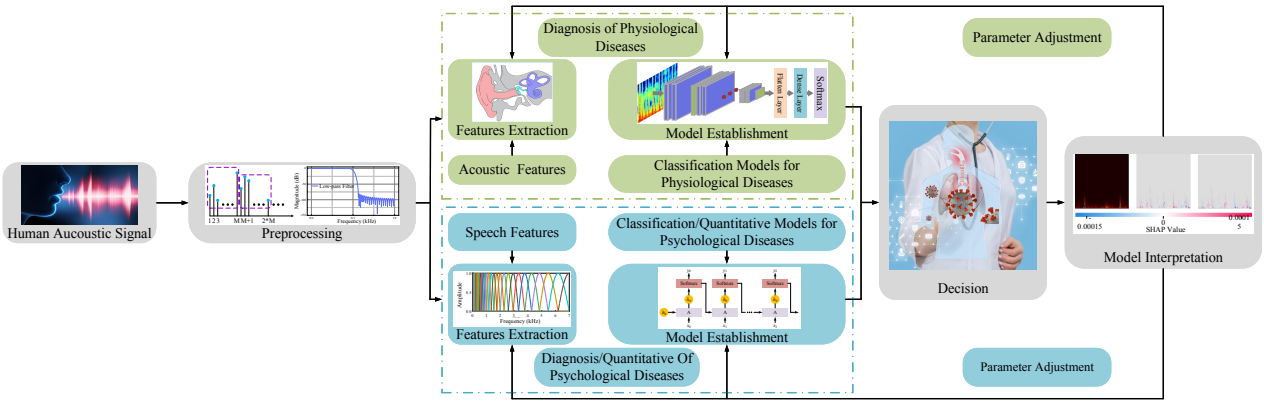

Fig. 1. Paradigm of CA4H. Audio can be used a "digital phenotype" [3] that represents the health status of the subject both physically and mentally.

Section 2 introduces the related work as complimentary to this introduction. Then, we illustrate the collected contributions in the first ICASSP Special Session on CA4H in Section 3. The current findings and limitations will be discussed in Section 4 before a conclusion is made in Section 5.

2. RELATED WORK

The early work on using audio to measure a subject's health status can be tracked back at least to the auscultation of bowel sound in the 1900's by Cannon [8]. Auscultation has been used an efficient method to understand what is happening in a human being's body for more than a century. Nevertheless, this professional skill is difficult for young trainees to acquire [9]. In the past two decades, tremendous efforts have been made towards using advanced signal processing (SP) and machine learning (ML) technologies to develop audio-enabled paradigms for diagnosis of diseases both physically and mentally [1]. Encouraging results were achieved in partiular in the analysis of snore sound [6], heart sound [10], lung sound [11], bowel sound [12], and joint sound [13]. Besides the aforementioned sounds generated by the body, speech can be used to diagnose manifold health states – e. g., autism spectrum disorder (ASD) [14], Alzheimer's disease (AD) [15], or major depressive disorder (MDD) [16]. However, CA is being under-featured in the community of digital health compared to CV. In this contribution, we highlight the state-of-the-art works presented at the first ICASSP Special Session on CA4H and indicate the open challenges that need to be addressed.

3. COLLECTED CONTRIBUTIONS

Table 1 illustrates the collected contributions in the session. In the following part of this section, we will introduce the collected work and their background separately.

3.1. COVID-19

At the time of writing this overview, COVID-19 has caused close to five million deaths (with more than 230 million confirmed cases) in a global wide [1]. Promising results have been found for COVID-19 detection methods by using cough and/or speech [23]. This is quite encouraging as an affordable, environmentally friendly, fast, and accurate diagnosis is important for control and further prevention of the spread of this disease [24]. Han et al. proposed a framework to detect COVID-19 from a subset of the audio sound data crowdsourced by their app [17]. Human hand-crafted features extracted by the OPENSMILE toolkit [25] and a 'classic' ML model – support vector machines (SVM) – were used. Four tasks were considered in their investigation: Positive vs negative, positive vs negative without symptoms, positive without symptoms vs negative without symptoms, and positive with symptoms vs negative with symptoms. They indicated that: First, the human voice is quite comparable to cough and breathing for COVID-19 diagnosis. Second, it is quite difficult to identify asymptomatic patients only from their voice. Third, when taking symptoms into account, the performance of the models could be improved.

3.2. Cough Detection

Liaqat et al. introduced the CoughWatach, a lightweight cough detector using audio and movement data in [18]. They used a large in-the-wild dataset for developing the models, and a smaller in-lab dataset for comparison to the existing work. A convolutional neural network (CNN) was used to extract high-level representations from the audio spectrograms generated by short-time Fourier transformation. They demonstrate that their proposed sensor fusion (SF) model can outperform other state-of-the-art works for both in-the-wild and in-lab datasets. In addition, they implemented the

[1] https://coronavirus.jhu.edu/map.html

Table 1. Collected contributions in the first ICASSP Special Session on Computer Audition for Healthcare (CA4H). AD: Alzheimer's disease. CVDs: Cardiovascular diseases (CVDs). AO: Audio Model. SF: Sensor Fusion. NNACE: Neural Networks based Advanced Combination Encoder. BiLSTM–RNN: Bidirectional Long Short-Term Memory Recurrent Neural Network. β–VAE: beta Variational Auto-Encoder. PCG: Phonocardiogram.

Ref.	Task	Methods	Remarks
[17]	COVID-19	OPENSMILE features SVM (*linear* kernel)	The voice signal shows great potential for developing an early-stage screening tool of COVID-19 (AUC: 0.79; Sensitivity: 68.0 %; Specificity: 82.0 %.).
[18]	Cough	AO Model SF Model	The proposed CoughWatch can achieve a precision of 82.0 % and recall of 55.0 % for cough detection. Incorporating gyroscope and accelerometer data is beneficial.
[19]	AD	Acoustic Model Linguistic Model	An early fusion of articulation and prosody features can reach the best performance (AUC: 0.77, 0.80). The linguistic models did not yield satisfactory results.
[20]	Cochlear	NNACE	The proposed NNACE can be compatible with the ACE-based CI systems. This framework considers both the noise reduction and the core envelope extraction.
[21]	Sress	OPENSMILE features BiLSTM–RNN	The stress can be detected via speech features. Adding contextual information (location and circadian rhythm) improves the baseline by 14.0 % in F1 scores.
[22]	CVDs	β–VAE	The β–VAE can be used to model the normal PCG signals in an unsupervised way. The beta value of the best model is smaller than one in most cases.

CoughWatch app on three different Android Wear smartwatches, which validated the feasibility of their method in terms of battery life and computation requirements.

3.3. Alzheimer's Disease

Pérez-Toro et al. investigated two models, i. e., acoustic and linguistic models for distinguishing the genetic Alzheimer's disease (AD) as well as early-onset Alzheimer's (EOA) related to the *Paisa mutation* [19]. For the acoustic model, articulation and prosody features were extracted for training the SVM model. For the linguistic model, the bidirectional encoder representations from a transformer were used. The acoustic model was found more efficient than the linguistic model in their study which might be due to the reason that the proportion of unknown words affect the model's performance.

3.4. Cochlear Implant

Zheng et al. proposed a neural network (NN) based advanced combination encoder (NNACE) for (cochlear implant) CI products of cochlear corporation [20]. In their strategy, a designed loss function was used for network training, which made the output signal noise-robust. Moreover, their method has low computational complexity. The authors conducted both objective and subjective evaluations to validate the success of their work. Their proposed method has utilised the NN model itself as a de-noising module such that no extra de-noising phase should be considered.

3.5. Stress-Level Monitoring

Real-time monitoring the stress of hospital workers is quite important to guarantee a high-quality service for patients, particularly during the COVID-19 pandemic. Gaballah et al. introduced the context-aware speech-based system for stress detection in [21]. For training models, human hand-crafted acoustic features and a bidirectional long short-term memory recurrent neural network (BiLSTM–RNN) were used. Their interesting finding was that the final performance can be improved by integrating location and circadian rhythm contextual cues along with the audio features.

3.6. Cardiovascular Disease Recognition

Cadiovascular diseases (CVDs) are ranked as the leading cause for deaths [26]. Li et al. presented their unsupervised learning method on automatically analysing the heart sound from the Phonocardiogram (PCG) [22]. In their study, a beta variational auto-encoder (β–VAE) was used to model the normal PCG signals. They indicated that the resampling process can help improve the anomaly PCG detection.

4. DISCUSSION

First of all, *data scarcity* is a serious challenge that can never be ignored. Similar to other data-driven applications in medicine, (accurately annotated) data is almost always rare, expensive, and difficult to share in the field of CA4H. Unlike most cases in speech resources, healthcare related audio data usually needs expert's knowledge to annotate. How to reduce the human expert's annotation work, and at the same time,

to guarantee an efficient performance of the models is a major concern in future CA4H studies. *Generative adversarial networks* (GANs) [27] have been found efficient in snore sound classification for data augmentation [28], which may be worth exploring for other diseases in future work.

Second, audio has a non-invasive characteristic by nature, which makes it quite suitable for its application both in the hospital and in home scenarios. However, only using audio may not yield the best performance [18]. Therefore, one should carefully consider the fusion of multi-modal data which may contribute to improve the model's performance only trained by audio.

Third, explainable AI (XAI) [29] is specifically important for medical applications. We may find that human handcrafted features (e. g., MFCCs, formants) can have clear physical meaning that can benefit exploring the relationship between the pathological mechanism and the acoustical properties. However, the high-level representations extracted via DL models may restrain a models' interpretability due to its "black–box" characteristic. In future work, building trustable and interpretable CA systems for medical applications should be taken into account.

Last but not the least, treatment might be more significant than diagnosis. When looking at the existing studies, most of them are aiming for leveraging audio data to assist the diagnosis progress. In contrast, finding novel treatment methods (e. g., non-drug induced ways) is crucial for both medical experts and patients. Music therapy [30] and dialog systems for therapy could be promising areas, in which more quantitative studies should be involved.

5. CONCLUSION

In this overview article, we summarised the collected contributions presented at the first ICASSP Special Session on Computer Audition for Healthcare (CA4H). On the one hand, we have witnessed encouraging results ranging from physiological disorders (e. g., hearing assistance, COVID-19 screening) to mental healthcare (e. g., stress measurement). On the other hand, we have to face the existing challenges among the current circumstance in this young community. We hope CA4H can be continuously serving as an open forum for colleagues from both academia and industry to work together towards the upcoming era of Medicine 4.0 by giving an emphasis on computer audition.

6. REFERENCES

[1] Kun Qian, Xiao Li, Haifeng Li, Shengchen Li, Wei Li, Zuoliang Ning, Shuai Yu, Limin Hou, Gang Tang, Jing Lu, et al., "Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, no. 5, pp. 1–4, 2020.

[2] Kun Qian, Zixing Zhang, Yoshiharu Yamamoto, and Björn W. Schuller, "Artificial intelligence internet of things for the elderly: From assisted living to healthcare monitoring," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 78–88, 2021.

[3] Sachin H Jain, Brian W Powers, Jared B Hawkins, and John S Brownstein, "The digital phenotype," *Nature biotechnology*, vol. 33, no. 5, pp. 462–463, 2015.

[4] Dirk Pevernagie, Ronald M Aarts, and Micheline De Meyer, "The acoustics of snoring," *Sleep Medicine Reviews*, vol. 14, no. 2, pp. 131–144, 2010.

[5] Patrick J Strollo Jr and Robert M Rogers, "Obstructive sleep apnea," *New England Journal of Medicine*, vol. 334, no. 2, pp. 99–104, 1996.

[6] Kun Qian, Christoph Janott, Maximilian Schmitt, Zixing Zhang, Clemens Heiser, Werner Hemmert, Yoshiharu Yamamoto, and Björn W. Schuller, "Can machine learning assist locating the excitation of snore sound? A review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1233–1246, 2021.

[7] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[8] Walter Bradford Cannon, "Auscultation of the rhythmic sounds produced by the stomach and intestines," *American Journal of Physiology-Legacy Content*, vol. 14, no. 4, pp. 339–353, 1905.

[9] Salvatore Mangione, "Cardiac auscultatory skills of physicians-in-training: A comparison of three english-speaking countries," *The American Journal of Medicine*, vol. 110, no. 3, pp. 210–216, 2001.

[10] Shahid Ismail, Imran Siddiqi, and Usman Akram, "Localization and classification of heart beats in phonocardiography signals—A comprehensive review," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 26, pp. 1–27, 2018.

[11] Rajkumar Palaniappan, Kenneth Sundaraj, and Nizam Uddin Ahamed, "Machine learning in lung sound analysis: A systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013.

[12] Gary Allwood, Xuhao Du, K Mary Webberley, Adam Osseiran, and Barry James Marshall, "Advances in acoustic signal processing techniques for enhanced bowel sound analysis," *IEEE Reviews in Biomedical Engineering*, vol. 12, pp. 240–253, 2019.

[13] Caitlin N Teague, Sinan Hersek, Hakan Töreyin, Mindy L Millard-Stafford, Michael L Jones, Géza F Kogler, Michael N Sawka, and Omer T Inan, "Novel methods for sensing acoustical emissions from the knee for wearable joint health assessment," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1581–1590, 2016.

[14] Kate Broome, Patricia McCabe, Kimberley Docking, and Maree Doble, "A systematic review of speech assessments for children with autism spectrum disorder: Recommendations for best practice," *American Journal of Speech-Language Pathology*, vol. 26, no. 3, pp. 1011–1029, 2017.

[15] María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert systems with applications*, vol. 150, pp. 1–19, 2020.

[16] Andrea Carolina Trevino, Thomas Francis Quatieri, and Nicolas Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[17] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo, "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8328–8332.

[18] Daniyal Liaqat, Salaar Liaqat, Jun Lin Chen, Tina Sedaghat, Moshe Gabel, Frank Rudzicz, and Eyal de Lara, "Coughwatch: Real-world cough detection using smartwatches," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8333–8337.

[19] PA Pérez-Toro, JC Vásquez-Correa, T Arias-Vergara, P Klumpp, M Sierra-Castrillón, ME Roldán-López, D Aguillón, L Hincapié-Henao, CA Tóbon-Quintero, T Bocklet, et al., "Acoustic and linguistic analyses to assess early-onset and genetic Alzheimer's disease," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8338–8342.

[20] Nengheng Zheng, Yupeng Shi, Yuyong Kang, and Qinglin Meng, "A noise-robust signal processing strategy for cochlear implants using neural networks," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8343–8347.

[21] Amr Gaballah, Abhishek Tiwari, Shrikanth Narayanan, and Tiago H Falk, "Context-aware speech stress detection in hospital workers using Bi-LSTM classifiers," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8348–8352.

[22] Shengchen Li, Ke Tian, and Rui Wang, "Unsupervised heart abnormality detection based on phonocardiogram analysis with beta variational auto-encoders," in *Proc. ICASSP*, Toronto, Canada, 2021, IEEE, pp. 8353–8357.

[23] Kun Qian, Björn W. Schuller, and Yoshiharu Yamamoto, "Recent advances in computer audition for diagnosing COVID-19: An overview," in *Proc. LifeTech*, Nara, Japan, 2021, pp. 185–186.

[24] Nastaran Taleghani and Fariborz Taghipour, "Diagnosis of COVID-19 for controlling the pandemic: A review of the state-of-the-art," *Biosensors and Bioelectronics*, vol. 174, pp. 112830: 1–17, 2020.

[25] Florian Eyben, Felix Weninger, Florian Gross, and Björn W. Schuller, "Recent developments in openSMILE, the Mmunich open-source multimedia feature extractor," in *Proc. ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[26] Elizabeth Wilkins, L Wilson, Kremlin Wickramasinghe, Prachi Bhatnagar, Jose Leal, Ramon Luengo-Fernandez, R Burns, Mike Rayner, and Nick Townsend, *European Cardiovascular Disease Statistics 2017*, European Heart Network, Brussels, Belgium, 2017.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, Montréal, Canada, 2014, pp. 2672–2680.

[28] Zixing Zhang, Jing Han, Kun Qian, Christoph Janott, Yanan Guo, and Björn W Schuller, "Snore-GANs: Improving automatic snore sound classification with synthesized data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, 2020.

[29] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[30] Nicole D Hahna, Susan Hadley, Vern H Miller, and Michelle Bonaventura, "Music technology usage in music therapy: A survey of practice," *The Arts in Psychotherapy*, vol. 39, no. 5, pp. 456–464, 2012.