AI Hears Your Health: Computer Audition for Health Monitoring

Shahin Amiriparian[1] and Björn Schuller[1,2](✉)

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
{shahin.amiriparian,bjoern.schuller}@uni-a.de
[2] GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK

Abstract. Acoustic sounds produced by the human body reflect changes in our mental, physiological, and pathological states. A deep analysis of such audio that are of complex nature can give insight about imminent or existing health issues. For automatic processing and understanding of such data, sophisticated machine learning approaches are needed that can extract or learn robust features. In this paper, we introduce a set of machine learning toolkits both for supervised feature extraction and unsupervised representation learning from audio health data. We analyse the application of deep neural networks (DNNs), including end-to-end learning, recurrent autoencoders, and transfer learning for speech and body-acoustics health monitoring and provide state-of-the-art results for each area. As show-case examples, we pick three well-benchmarked examples for body-acoustics and speech, each, from the popular annual Interspeech Computational Paralinguistics Challenge (ComParE). In particular, the speech-based health tasks are COVID-19 speech analysis, recognition of upper respiratory tract infections, and continuous sleepiness recognition. The body-acoustics health tasks are COVID-19 cough analysis, speech breath monitoring, heartbeat abnormality recognition, and snore sound classification. The results for all tasks demonstrate the suitability of deep computer audition approaches for health monitoring and automatic audio-based early diagnosis of health issues.

Keywords: Computer audition · Digital health · Health monitoring

1 Introduction

Diagnosis of disease, ideally even before symptoms are noticeable to individuals, facilitates early interventions and maximises the chance of successful treatments, especially for mental health. Whilst early diagnosis cannot enable curative treatment of all possible diseases, it provides the considerable chance of averting irreversible pathological changes in organ, skeletal, and nervous systems, as well as

E. Pissaloux et al. (Eds.): IHAW 2021, CCIS 1538, pp. 227–233, 2021.
https://doi.org/10.1007/978-3-030-94209-0_20

chronic pain and psychological stress [8]. Research in machine learning for audio-based digital health applications has increased in recent years [6]. Substantial contributions have been made to the development of audio-based techniques for the recognition of various health conditions, including neurodegenerative diseases such as Alzheimer's or Parkinson's [20], psychological disorders such as bipolar disorder [16], neurodevelopmental disorders such as Fragile X, Rett-Syndrome, or Autism Spectrum Disorder [17], and contagious diseases such as COVID-19 [15]. In the proceeding section of this paper, we first introduce seven health-related corpora for speech and acoustic health monitoring tasks (Sect. 2). In Sect. 3, we then introduce a set of contemporary computer audition methods and analyse their performance for various early digital health diagnosis and recognition tasks. The last section concludes our paper and discusses future work.

2 Speech and Acoustic Health Datasets

In this section, we introduce seven health related speech and audio datasets which have been used in recent editions of the INTERSPEECH Computational Paralinguistics ChallengE (COMPARE) [18,19,22]. We further provide information about the important characteristics of each dataset and the used partitions for the machine learning experiments (cf. Table 1).

Cambridge COVID19 Sound Database – Speech & Cough. This dataset which was used for a sub-challenge in the 2019 edition of the INTERSPEECH ComParE contains two speech and cough subsets from the Cambridge COVID-19 Sound database [3,11]. The audio files were resampled (in some cases, upsampled) and then converted to 16 kHz and mono/16 bit, and further normalised recording-wise to eliminate varying loudness. For the COVID-19 Cough (C19C), 725 recordings (one to three forced coughs) from 343 participants were provided, in total 1.63 h. For the COVID-19 Speech (C19S), 893 speech recordings from 366 individuals were used, in total 3.24 h.

Upper Respiratory Tract Infection Corpus (URTIC). This corpus is provided by the Institute of Safety Technology, University of Wuppertal, Germany, and consists of recordings of 630 subjects (382 m, 248 f, mean age 29.5 years, std. dev. 12.1 years, range 12-84 years), made in quiet rooms with a microphone/headset/hardware setup (sample rate 44.1 kHz, downsampled to 16 kHz, quantisation 16 bit). To obtain the state of health, each individual reported a binary one-item measure based on the German version of the Wisconsin Upper Respiratory Symptom Survey (WURSS-24), assessing the symptoms of common cold. The global illness severity item (on a scale of 0 = not sick to 7 = severely sick) was binarised using a threshold at 6.

Düsseldorf Sleepy Language (SLEEP) Corpus. This corpus [21] contains speech recordings of 915 individuals (364 f, 551 m) at different levels of sleepiness (1–9 KSS, 9 denotes extreme sleepiness). The participants performed various pre-defined speaking tasks and read out text passages. Moreover, spontaneous speech is collected in the form of elicited narrative content. The sessions which

Table 1. Number of instances per class in the all partitions for each dataset.

#	Training	Development	Test	Σ
Speech-based datasets for health monitoring				
COVID-19 Speech (C19S) Corpus [3,11,23]				
No COVID-19	243	153	189	585
COVID-19	72	142	94	308
Σ	315	295	283	893
Upper Respiratory Tract Infection Corpus (URTIC) [19]				
C	970	1 011	895	2 876
NC	8 535	8585	8 656	25 776
Σ	9 505	9 596	9 551	28 652
Düsseldorf Sleepy Language (SLEEP) Corpus [21]				
1–9 (Karolinska Sleepiness Scale (KSS))	5 564	5 328	5 570	16 462
Acoustic datasets for health monitoring				
COVID-19 Cough (C19C) Corpus [3,11,23]				
No COVID-19	215	183	169	567
COVID-19	71	48	39	158
Σ	286	231	208	725
UCL Speech Breath Monitoring (UCL-SBM) Corpus [18]				
Speakers	17	16	16	49
Heart Sounds Shenzhen (HSS) Corpus [22]				
Normal	84	32	28	144
Mild	276	98	91	465
Moderate/Severe	142	50	44	236
Σ	502	180	163	845
Munich-Passau Snore Sound Corpus (MPSSC) [19]				
Velum (V)	168	161	155	484
Oropharyngeal lateral walls (O)	76	75	65	216
Tongue (T)	8	15	16	39
Epiglottis (E)	30	32	27	89
Σ	282	283	263	828

lasted roughly one hour per participant were further held between 6 am to 12 pm in order to acquire high variability in the levels of perceived sleepiness. Using this dataset, the sleepiness of a speaker can be assessed as regression problem. Continuous recognition of sleepiness is of high relevance for sleep disorder monitoring.

UCL Speech Breath Monitoring (UCL-SBM) Corpus. This corpus contains spontaneous speech recordings that took place in a quiet office space, and recordings from a piezoelectric respiratory belts worn by the subjects. All signals were sampled at 40 kHz; speech was downsampled to 16 kHz and breath belts to 25 Hz in post-processing [18]. All 49 speakers (29 f, 20 m) reported English as a primary language ages range from 18 to approximately 55 years old (mean age 24 years; std. dev. ~10 years). Breathing patterns also provide medical doctors vital information about an individual's respiratory and speech planning [4].

Heart Sounds Shenzhen (HSS) Corpus. The HSS corpus, provided by the Shenzhen University General Hospital, contains heart sounds gathered from 170 subjects (55 f, 115 m; ages from 21 to 88 years (mean age 65.4 years, std. dev. 13.2 years) with various health conditions, such as coronary heart disease, heart failure, and arrhythmia. The acoustic signals were recorded using an electronic stethoscope with a 4 kHz sampling rate and a 20 Hz–2 kHz frequency response. Three types of heartbeats (normal, mild, and moderate/severe) have to be classified Table 1. Automatic machine learning based approaches could help monitoring patients with unclear symptoms of heartbeat abnormalities.

Munich-Passau Snore Sound Corpus (MPSSC). The MPSSC is introduced for classification of snore sounds by their excitation location within the upper airways. The corpus contains audio samples of 828 snore events from 219 subjects (cf. Table 1). The number of recordings per class in the corpus is unbalanced, with 84% of samples from the classes Velum (V) and Oropharyngeal lateral walls (O), 11%, Epiglottis (E)-events, and 5% Tongue (T)-snores. This is in line with the probability of occurrence during normal sleep [12].

Table 2. Results for all seven introduced corpora. The **official challenge baselines** and the winners of each sub-challenge are provided. UAR: Unweighted Average Recall. PCC: Pearson's correlation coefficient. ρ: Spearman's correlation coefficient. *: [2] was a separate submission and not as a part of the sub-challenge.

	Speech-based health monitoring						**Acoustic health monitoring**							
Approach	**C19S** UAR [%]		**URTIC** UAR [%]		**SLEEP** PCC		**C19C** ρ		**UCL-SBM** UAR [%]		**HSS** UAR [%]		**MPSSC** UAR [%]	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Baseline systems of the ComParE [23, 19, 21, 22, 2]														
openSMILE	57.9	**72.1**	64.0	70.2	.251	.314	61.4	65.5	.244	.442	50.3	46.4	40.6	58.5
End2You	70.5	68.8	59.1	60.0	N/A		61.8	64.7	.507	.731	41.2	37.7	40.3	40.3
auDeep	62.2	64.2	N/A		.257	.321	67.6	67.6	N/A		38.6	47.9	44.8	61.3
Deep Spectrum	56.0	60.4	N/A		N/A		63.3	64.1	N/A		44.1	46.1	44.8	67.0*
Fusion of Best	–	71.1	–	71.0	–	.343	–	73.9	–	.621	–	**56.2**	–	55.8
Winners of each sub-challenge from left to right: [10, 14, 9, 5, 13]														
	baseline won		**65.8**	**72.0**	**.367**	**.383**	**69.9**	**75.9**	**.640**	**.763**	**baseline won**		–	**64.2**

3 State-of-the-Art Methodologies and Results

This section provides results from the winners of each sub-challenge (cf. Table 2). Further, the results are compared with the performance of four machine learning and deep learning baseline systems of ComParE, namely openSMILE[1] [7], End2You[2] [24], auDeep[3] [1], and Deep Spectrum[4] [2]. Each of baseline system utilises a different methodology to extract or learn features from the audio signals. In particular, openSMILE is designed to extract expert-designed features such as pitch, energy, and prosody for specific speech and audio tasks. The End2You approach utilises an end-to-end learning paradigm to extract features from raw audio with a convolutional network and then performing the final classification using a subsequent recurrent network. auDeep makes use of recurrent sequence-to-sequence autoencoders for unsupervised representation learning, and Deep Spectrum applies transfer learning techniques with pre-trained image convolutional networks for deep feature extraction from audio plots.

4 Conclusions and Future Work

We have carefully selected seven (three speech-based and three body-acoustics-based plus one 'inbetweener' – breathing) medical datasets for audio-based early diagnosis of various health issues (cf. Sect. 2), and demonstrated the suitability of (deep) computer audition methods for all introduced tasks (cf. Sect. 3). For data of a more complex nature (e. g. SLEEP or C19C), we showed that unsupervised learning of representations provides better results compared to other baselines. For the regression task UCL-SBM, End2You (composed of convolutional and recurrent blocks) outperforms other systems showing its suitability for modelling time-continuous data. Further, we recommend the application of transfer learning approaches (e. g. Deep Spectrum) for audio health monitoring tasks where the data is scarce as such models are pre-trained on larger datasets. As a next step, more holistic views on audio-based health monitoring will be needed that do not focus on 'healthy' vs 'sick', but target the big picture of health state synergistically. With this and more data or data-efficient strategies, audio-based health monitoring in every-day life appears around the corner.

References

1. Amiriparian, S., Freitag, M., Cummins, N., Schuller, B.: Sequence to sequence autoencoders for unsupervised representation learning from audio. In: Proceedings of DCASE 2017, Munich, Germany, pp. 17–21 (2017)

[1] https://github.com/audeering/opensmile.
[2] https://github.com/end2you/end2you.
[3] https://github.com/auDeep/auDeep.
[4] https://github.com/DeepSpectrum/DeepSpectrum.

2. Amiriparian, S., et al.: Snore sound classification using image-based deep spectrum features. In: Proceedings of Interspeech 2017, Stockholm, Sweden, pp. 3512–3516 (2017)
3. Brown, C., Chauhan, J., Grammenos, A., et al.: Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: Proceedings of KDD, San Diego, CA, pp. 3474–3484 (2020)
4. Capellan, A., Fuchs, S.: The interplay of linguistic structure and breathing in German spontaneous speech. In: Proceedings of Interspeech, Lyon, France (2013)
5. Casanova, E., Candido Jr., A., Fernandes Jr., R.C., et al.: Transfer learning and data augmentation techniques to the COVID-19 identification tasks in ComParE 2021. In: Proceedings of Interspeech 2021, pp. 446–450 (2021)
6. Deshpande, G., Schuller, B.: An overview on audio, signal, speech, & language processing for COVID-19. arXiv preprint arXiv:2005.08579 (2020)
7. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia, pp. 1459–1462. ACM (2010)
8. Fufurin, I.L., Golyak, I.S., Anfimov, D.R., et al.: Machine learning applications for spectral analysis of human exhaled breath for early diagnosis of diseases. In: Optics in Health Care and Biomedical Optics X, vol. 11553, p. 115531G. International Society for Optics and Photonics (2020)
9. Gosztolya, G.: Using fisher vector and bag-of-audio-words representations to identify styrian dialects, sleepiness, baby & orca sounds (2019)
10. Gosztolya, G., et al.: DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification (2017)
11. Han, J., Brown, C., Chauhan, J., et al.: Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In: Proceedings of ICASSP, Toronto, Canada (2021)
12. Hessel, N.S., de Vries, N.: Diagnostic work-up of socially unacceptable snoring. Eur. Arch. Otorhinolaryngol. **259**(3), 158–161 (2002). https://doi.org/10.1007/s00405-001-0428-8
13. Kaya, H., Karpov, A.A.: Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: snoring, addressee and cold. In: INTERSPEECH, pp. 3527–3531 (2017)
14. Markitantov, M., Dresvyanskiy, D., Mamontov, D., et al.: Ensembling end-to-end deep models for computational paralinguistics tasks: compare 2020 mask and breathing sub-challenges. In: INTERSPEECH, pp. 2072–2076 (2020)
15. Qian, K., Schuller, B.W., Yamamoto, Y.: Recent advances in computer audition for diagnosing COVID-19: an overview. In: 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), pp. 181–182. IEEE (2021)
16. Ringeval, F., Schuller, B., Valstar, et al.: AVEC 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition. In: Proceedings of the 2018 on Audio/visual Emotion Challenge and Workshop, pp. 3–13 (2018)
17. Roche, L., Zhang, D., Bartl-Pokorny, K.D., et al.: Early vocal development in autism spectrum disorder, rett syndrome, and fragile x syndrome: insights from studies using retrospective video analysis. Adv. Neurodevelop. Disorders **2**(1), 49–61 (2018). https://doi.org/10.1007/s41252-017-0051-3
18. Schuller, B., Batliner, A., Bergler, C., et al.: The interspeech 2020 computational paralinguistics challenge: elderly emotion, breathing & masks. In: Proceedings INTERSPEECH 2020, ISCA, pp. 2042–2046 (2020)

19. Schuller, B., Steidl, S., Batliner, A., Bergelson, et al.: The interspeech 2017 computational paralinguistics challenge: addressee, cold & snoring. In: Proceedings INTERSPEECH 2017, pp. 3442–3446 (2017)
20. Schuller, B., Steidl, S., Batliner, A., et al.: The INTERSPEECH 2015 computational paralinguistics challenge: degree of nativeness, Parkinson's & eating condition. In: Proceedings of Interspeech, Dresden, Germany, pp. 478–482 (2015)
21. Schuller, B.W., Batliner, A., Bergler, C., et al.: The INTERSPEECH 2019 computational paralinguistics challenge: styrian dialects, continuous sleepiness, baby sounds & orca activity. In: Proceedings INTERSPEECH 2019, ISCA, ISCA, Graz, Austria, pp. 2378–2382 (2019)
22. Schuller, B.W., et al.: The INTERSPEECH 2018 computational paralinguistics challenge: atypical & self-assessed affect, crying & heart beats. In: Proceedings of INTERSPEECH 2018, pp. 122–126 (2018)
23. Schuller, B.W., Batliner, A., Bergler, C., et al.: The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. In: Proceedings INTERSPEECH 2021, ISCA, Brno, Czechia (2021)
24. Tzirakis, P., Zafeiriou, S., Schuller, B.W.: End2You-the imperial toolkit for multimodal profiling by end-to-end learning. arXiv preprint arXiv:1802.01115 (2018)