2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Transferring Cross-Corpus Knowledge: An Investigation on Data Augmentation for Heart Sound Classification

Tomoya Koike[1], *Student Member, IEEE*, Kun Qian[1,2], *Senior Member, IEEE*,
Björn W. Schuller[3,4], *Fellow, IEEE*, and Yoshiharu Yamamoto[1], *Member, IEEE*

***Abstract*— Human auscultation has been regarded as a cheap, convenient and efficient method for the diagnosis of cardiovascular diseases. Nevertheless, training professional auscultation skills needs tremendous efforts and is time-consuming. Computer audition (CA) that leverages the power of advanced machine learning and signal processing technologies has increasingly attracted contributions to the field of automatic heart sound classification. While previous studies have shown promising results in CA based heart sound classification with the 'shuffle split' method, machine learning for heart sound classification decreases in accuracy with a cross-corpus test dataset. We investigate this problem with a cross-corpus evaluation using the PhysioNet CinC Challenge 2016 Dataset and propose a new combination of data augmentation techniques that leads to a CNN robust for such cross-corpus evaluation. Compared with the baseline, which is given without augmentation, our data augmentation techniques combined improve by 20.0 % the sensitivity and by 7.9 % the specificity on average across 6 databases, which is a significant difference on 4 out of these ($p < .05$ by one-tailed z-test).**

I. Introduction

Computer audition (CA) has been increasingly applied to the field of healthcare [1], e. g., for sleep disorder [2], and in the ongoing COVID-19 pandemic [3], [4]. In particular, CA-based methods can facilitate a possible alternative to human auscultation. Auscultation is regarded as a cheap, convenient and efficient method for diagnosis of cardiovascular diseases (CVDs), the leading cause of human deaths [5].

To automatically and correctly classify abnormal heart sounds from normal ones, a lot of machine learning techniques are proposed in the previous literature [6]. Short-time Fourier transform (STFT) and continuous wavelet transform (CWT) are used to create handcrafted features with a variety of classifiers which includes support vector machines (SVM), K-nearest neighbours (KNN), and artificial neural networks (ANN) [7]. Without handcrafted features, feature distribution leaning from the raw data or a 'simple' spectrogram with convolutional neural networks (CNNs) is focused on these days [8], [9], achieving high accuracy on the PhysioNet CinC Challenge 2016 database (CinC DB) [10].

Pre-training on a large dataset increases the accuracy in many tasks, e. g., in computer vision [11], natural language processing [12], and speech recognition [13]. Also in heart sound classification task, CNNs with pre-training on other large datasets such as ImageNet [14] and AudioSet [15] usually boosts the model accuracy by transfer learning [16], [17]. Although CNNs have the large capability of fitting data distributions with many learnable parameters; yet, they tend to overfit and worsen on a test dataset particularly if the test dataset is not similar to the training dataset. The test dataset is chosen randomly in [17] and VGG [18] scored 93.7 % for sensitivity on the CinC DB. In [16], the test dataset is chosen from two databases within CinC DB coming from two different hospitals, and training is conducted on other databases except those two, and VGG scored 33.3 % for sensitivity.

Though different hyperparameters and training methods, it is highly possible that a CNN overfits to its training dataset and performance deteriorates on a highly different-from-training test dataset. Cross-corpus evaluation is mimicring the situation of a real-world application for heart sound classification in a 'new' hospital, as 6 databases included in CinC DB are collected from different hospitals. Regardless of many previous works for heart sound classification, it is only a few of them which deploy proper 'realistic' evaluation methods for a real-world application and propose accordingly robust learning schemes.

To tackle cross-corpus heart sound classification, we introduce new data augmentation techniques with a proper cross-corpus evaluation approach. In addition to audio data augmentations previously proposed, random trimming and the effects of respiration on heart sound are reproduced and evaluated in this paper. By excluding one database as a cross-corpus test dataset from a collection of different recording sites, a real-world situation is reproduced using 6 databases within the CinC DB collection and overfitting of the CNN learning model is revealed. We show that the this overfitting of the CNN is attenuated by our data augmentation technique and scores on the cross-corpus datasets are improved.

The rest of this work is structured as follows: Data

[1] Tomoya Koike, Kun Qian, and Yoshiharu Yamamoto are with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. {tommy, qian, yamamoto}@p.u-tokyo.ac.jp

[2] Kun Qian is also with the Group on Audition for Intelligent Medicine (AIM), Institute of Engineering Medicine, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing 100081, China. qiantum@hotmail.com

[3,4] Björn W. Schuller is with GLAM – the Group on Language, Audio & Music, Imperial College London, 180 Queens' Gate, Huxley Bldg., London SW7 2AZ, UK, and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Eichleitnerstr. 30, Augsburg 86159, Germany. schuller@ieee.org

This work was partially supported by the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, and the Grants-in-Aid for Scientific Research (No. 19F19081 and No. 20H00569) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. Kun Qian is the *Corresponding author.*

augmentation and the transfer learning model are introduced in Section II. The experimental setup and results are shown in Section III, followed by discussion part in Section IV.

II. Materials and methods

A. Audio data augmentation

Audio data augmentation has been proposed as a method to generate additional training data [19], [20]. SpecAugment [21] is masking on a spectrogram which has a shape of a line with some width in the time axis and/or the frequency axis. This reproduces a malfunctioning recording system with a time axis mask and machine noise with a frequency axis mask. In heart sound data, hitting something onto the microphone generates a pulse-like sound, which is similar to a time mask on a spectrogram. Medical machines generating a sound in some frequency bands appear in the spectrogram similar to the frequency mask of SpecAugment. Random amplitude change is to multiply a constant value to a sound and scale the amplitude at each sample point. This is observed, if the default setting value of the sensor is different, or the position of the sensor in a body is different. Random erase can be a reproduction of the situation that specific background noise is observed, or the information an input spectrogram contains is randomly dropped. This reproduction is conducted by masking an input spectrogram with a rectangular mask of a constant value.

B. Heart sound data augmentation

A heart sound signal, or more generally, a biological signal, has some specific characteristics compared with general audio sound, for example, a more periodic nature and a longer period.

Respiration is a major factor to affect heart sound, but is not related to heart sound abnormality. As you can see in Figure 1, some sounds are marked by fluctuation which is periodic at some period and contain background sounds similar to breathing. This situation can be reproduced by generating sinusoidal weights to be multiplied to an original sound at each sample point. As respiration is periodic, and a recording of a heart sound is irrelevant to a respiration period, the phase of sinusoidal weights should be randomly decided from 0 to $2 \times \pi$. The amount of the respiration effect on heart sound is not clear, and the respiration rate, which corresponds to the frequency of sinusoidal weights, depends on the subject. Due to the beats' periodicity, normal as well as abnormal sound is repeated in a recording and not found at a constant position in a spectrogram, which leads to the idea of random trimming with a constant length. By deciding the length of trimming – 5 seconds in this paper – a sound can be augmented from one long recording into many short segments. It should be noted that non-periodic abnormal heart sound such as arrhythmia can be ignored with this augmentation.

C. Transfer learning model

Many CNNs have by now been proposed for the heart sound classification task. Such CNN allow for pre-training on other large datasets to boost the accuracy of the model. The ImageNet dataset is a popular choice for pre-training in the audio domain, but the AudioSet dataset is an alternative choice for sound data, which contains a lot of recordings for sound event classification and hence appears more suited. Based on this data, pre-trained audio neural networks (PANNs) [22] are introduced here for the task at hand, as they achieved state-of-the-art on some audio pattern recognition tasks by pre-training the model on the AudioSet dataset. CNN14, which is one of the CNN structures proposed in PANNs and the one we use in this paper, has 14 convolutional layers and other layers in total to extract features from a spectrogram. The weights in the network after pre-training on the AudioSet dataset are publicly available. We initialise CNN14 with this weight and start training on CinC DB. The number of output nodes of the last fully connected layer is changed to two to suit the heart sound classification task.

TABLE I
Overview of the 6 databases included in CinC DB.

Database	Recordings	Normal	Abnormal	Duration [s]
MIT	409	117	292	32.6
AAD	490	386	104	7.9
AUTH	31	7	24	49.4
UHA	55	27	28	15.1
DLUT	2 141	1 958	183	23.1
SUA	114	80	34	33.1

III. Experimental Results

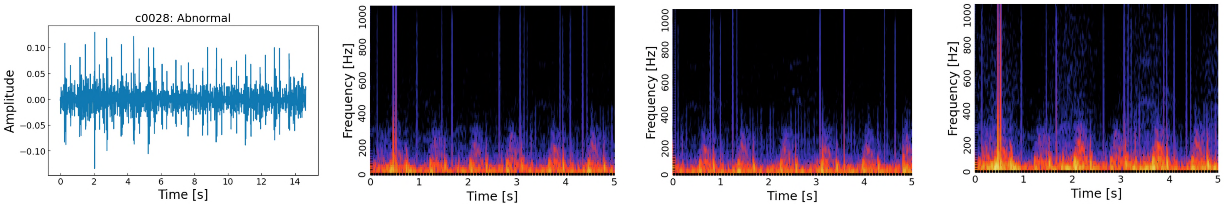
A. Dataset

To evaluate each data augmentation approach and transfer learning model, as mentioned, we use six databases included in CinC DB which is publicly distributed[1]. Each database has different characteristics in terms of the number of recordings, the number of labels in each class, and the mean length of the sound, which are described in detail in Table I. The shortest length of the recordings is 5.3 seconds in the "AAD" database and the longest is 122.0 seconds in the "AUTH" database. It should be noted that some databases have a small number of recordings, and this can be a problem when that database is chosen as a test dataset. We avoid this problem by averaging the results with different seeds in the experiment.

B. Pre-processing

The raw heart sound is at first trimmed into a 5 second clip by extracting the middle of a sound and then transformed into a spectrogram, which results in a 2-dimensional feature map with time and frequency axis. Standardisation of the spectrogram is not primarily considered as a data augmentation, but a simple approach to ease the difference of the data distribution by subtracting mean and dividing by the standard deviation. Random amplitude change, random trimming, and respiratory scaling are applied on the raw waveform; frequency masking, time masking, standardisation, and random erase are applied on the spectrogram. The 5 second clips of heart sounds are transformed into a spectrogram with a 100 milliseconds

[1] https://physionet.org/content/challenge-2016/1.0.0/



(a) Raw waveform (b) STFT (c) Random trimming (d) Respiratory scaling

Fig. 1. Simple STFT and two data augmentation techniques for heart sound classification. Raw heart sound is trimmed to 5 seconds before calculating a spectrogram.

TABLE II

UAR [%] OF EACH TEST DATABASE (COLUMNS) WITH EACH DATA AUGMENTATION APPROACH (LINES). REPORTED NUMBERS ARE AVERAGES OVER FIVE DIFFERENT SEED VALUES. THE AUGMENTATION TECHNIQUE DENOTED AS "ALL" RESEMBLES APPLICATION OF ALL AUGMENTATION TECHNIQUES.

Augmentation	AAD	AUTH	DLUT	MIT	SUA	UHA
Baseline	58.6	64.5	49.3	55.4	50.6	57.7
Frequency mask	59.9	**76.0**	52.6	56.8	50.8	**66.1**
Time mask	59.3	75.8	49.2	**57.7**	50.6	60.0
Standardisation	60.3	71.4	55.8	56.9	52.3	59.2
Random amp.	59.3	68.5	52.1	55.0	51.7	60.0
Random erase	57.1	71.3	48.9	56.9	52.2	64.9
Random trim.	61.2	59.3	47.6	55.1	54.1	62.7
Resp. scaling	58.6	64.9	51.9	54.6	50.1	57.3
All	**62.1**	67.1	**60.2**	54.3	**55.3**	56.3

window, 25 milliseconds stride, and zero-padding to reach 200 milliseconds in each Fourier transform, generating 101 × 401 feature map. The CNN14 is trained with the following hyperparameters: the batch size is 32 with 30 epochs; the learning rate is 0.001 decreasing in each epoch by 0.99 times; the weights are optimised with Adam [23].

C. *Evaluation Method and Metrics*

The evaluation of the robustness is conducted by the leave-one-DB-out method, which chooses one database as a test dataset and uses the other databases to train a model. As stated, we conduct leave-one-DB-out with all 6 databases included in the CinC DB, and 5 different seed values to avoid outlying results due to a small number of instances in a test dataset. 5 of the 6 databases (leaving out the test dataset) are shuffled and split into a training dataset and a validation dataset to train the model and decide when to stop training to avoid overfitting. To deal with label imbalance in each database, we sample instances of a training dataset repeatedly to render the numbers of "normal" and "abnormal" instances equal when training the model.

Following official score metrics of the PhysioNet CinC Challenge 2016 [10], sensitivity and specificity are used as evaluation metrics.

Unweighted average recall (UAR) [24] is also used as an evaluation metric.

D. *Results*

The experimental results are shown in Table II and Table III (in [%]). The augmentation denoted as "All" represents the experimental condition in which all augmentation techniques introduced in Section II are deployed at once – each with a probability of 0.1. All of the results shown in Table II and Table III again reflect the average of 5 different seed values. The baseline results denoted as "Baseline" are 56.0 % UAR, 35.2 % sensitivity, and 60.2 % specificity as an average over the 6 test databases. The highest sensitivity and specificity are 51.6 % and 64.1 % – both with all augmentation methods combined, which resembles improvements over the baseline by 20.0 %, and 7.9 %, respectively. Of the 6 databases by UAR, 'All' augmentation applied reaches the maximum on 3 databases: the AAD, DLUT, and SUA databases. "All" augmentations combined further marks the best scores on the AUTH, MIT, and UHA databases both for sensitivity and specificity.

IV. Discussion

Looking further at the results as to the individual methods, random trimming reaches the second-best UAR, sensitivity, and specificity on the AAD database, and the second-best sensitivity on the UHA database. Compared with no augmentation, respiratory scaling provides the second-best sensitivity and specificity on the DLUT database, and the first best sensitivity on the SUA database. As we list some augmentation techniques and the scores in Section III-D, specific augmentation methods are the most effective on some databases. Frequency mask appears useful to achive high robustness on the AUTH and UHA databases, Effective data augmentation implies coping with how different train databases and test database are from each other, e. g., the AUTH and UHA databases contain some noise bands along the frequency axis. Considering the small number of instances in a test dataset, we conduct a one-tailed z-test on the results of 'All' augmentation on all of the 6 databases. The improvement is significant on AAD, AUTH, DLUT, and UHA databases ($p < .05$) for sensitivity, while it is only significant for the DLUT database ($p < .001$) regarding UAR and specificity. Assuming that real data from a hospital can be different from all of the databases included in a training dataset, the combination of all data augmentation methods appears promising to obtain improved sensitivity, as we reproduced those situations by CinC DB.

V. Conclusion

In this study, we investigated data augmentation techniques for cross-corpus heart sound classification. We listed popular

TABLE III

SENSITIVITY AND SPECIFICITY OF EACH TEST DATABASE WITH EACH DATA AUGMENTATION TECHNIQUE. REPORTED RESULTS ARE AVERAGES OVER FIVE DIFFERENT SEED VALUES.

Augmentation	Sensitivity							Specificity						
	AAD	AUTH	DLUT	MIT	SUA	UHA	Mean±Std.	AAD	AUTH	DLUT	MIT	SUA	UHA	Mean±Std.
Baseline	47.5	43.3	44.6	26.0	24.1	25.7	35.2±11.0	83.3	31.4	91.2	31.5	70.0	54.0	60.2±25.6
Freq. mask	58.1	63.3	45.8	30.6	35.3	49.3	47.1±12.7	84.8	42.4	92.3	32.6	69.7	62.0	64.0±23.3
Time mask	64.4	65.8	35.4	27.1	21.8	39.3	42.3±18.7	85.2	43.0	91.1	32.7	70.6	56.3	63.2±23.3
Standardise	52.3	60.0	85.9	26.1	35.3	24.3	47.3±23.7	83.9	39.0	95.2	32.3	68.5	54.7	62.3±24.9
Random amp.	52.9	48.3	46.0	20.5	25.9	37.9	38.6±13.0	83.7	34.9	92.0	31.2	72.5	56.7	61.9±25.3
Random erase	40.8	62.5	55.7	36.0	37.6	53.6	47.7±11.0	82.6	41.3	91.1	33.1	73.1	62.6	64.0±23.0
Random trim.	57.3	35.8	29.2	34.2	22.9	54.3	39.0±13.8	85.0	28.5	90.8	31.8	72.6	61.5	61.7±26.5
Resp. scaling	46.5	44.2	56.4	28.8	61.2	25.0	43.7±14.5	83.0	31.3	92.9	31.2	67.6	55.0	60.2±25.9
All	67.7	74.1	83.7	25.6	17.6	40.7	51.6±27.4	87.4	43.1	95.6	31.0	72.7	55.0	64.1±25.4

TABLE IV

CONFUSION MATRICES (NORMALISED: IN [%]) OF EACH TEST SET PREDICTED BY THE MODEL WITH THE COMBINATION OF AUGMENTATION METHODS. CONFUSION MATRICES ARE SUMMED AND NORMALISED OVER FIVE SEEDS. **N**: NORMAL; **A**: ABNORMAL.

(a) AAD

	N	**A**
N	56.6	43.4
A	32.3	67.7

(b) AUTH

	N	**A**
N	60.0	40.0
A	25.8	74.2

(c) DLUT

	N	**A**
N	36.6	63.4
A	16.3	83.7

(d) MIT

	N	**A**
N	82.9	17.9
A	74.4	25.6

(e) SUA

	N	**A**
N	93.0	7.0
A	82.4	17.6

(f) UHA

	N	**A**
N	71.9	28.1
A	59.3	40.7

audio augmentation techniques and proposed further ones, especially for heart sound data. To reproduce a real application situation, where collected heart sound data will be different in its characteristics from a training dataset, we evaluated a transfer CNN model leaving out entire databases for testing. We found that all augmentations combined improved by 20.0 % sensitivity and 7.9 % specificity, which resembles a significant difference on 4 out of 6 databases ($p < .05$ by one-tailed z-test), compared to no augmentation. Concluding, all augmentation techniques combined are promising for cross-corpus heart sound classification in terms of sensitivity and encourage future consideration of further such.

REFERENCES

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

[2] K. Qian *et al.*, "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.

[3] K. Qian, B. W. Schuller, and Y. Yamamoto, "Recent advances in computer audition for diagnosing COVID-19: An overview," in *Proc. LifeTech*. Nara, Japan: IEEE, 2021, pp. 185–186.

[4] K. Qian *et al.*, "Computer audition for fighting the SARS-CoV-2 corona crisis – Introducing the multi-task speech corpus for COVID-19," *IEEE Internet of Things Journal*, pp. 1–12, 2021, in press.

[5] A. Timmis *et al.*, "European Society of Cardiology: Cardiovascular Disease Statistics 2017," *European Heart Journal*, vol. 39, no. 7, pp. 508–579, 11 2017.

[6] S. Ismail *et al.*, "Localization and classification of heart beats in phonocardiography signals–A comprehensive review," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 26, 2018.

[7] F. Dong *et al.*, "Machine listening for heart status monitoring: Introducing and benchmarking HSS–the heart sounds Shenzhen corpus," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2082–2092, 2020.

[8] J. Rubin *et al.*, "Recognizing abnormal heart sounds using deep learning," in *Proc. KHD Workshop of the IJCAI*, Melbourne, Australia, 2017, pp. 13–19.

[9] C. Potes *et al.*, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proc. CinC*, Columbia, Canada, 2016, pp. 621–624.

[10] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, 2016.

[11] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 2661–2671.

[12] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing." *arXiv preprint arXiv:1910.07370*, pp. 1–11, 2019.

[13] C.-X. Qin, D. Qu, and L.-H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–9, 2018.

[14] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, USA, 2009, pp. 248–255.

[15] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] Z. Ren *et al.*, "Learning image-based representations for heart sound classification," in *Proc. DH*. Lyon, France: ACM, 2018, p. 143–147.

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

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

[19] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. ASRU*, 2013, pp. 309–314.

[20] A. Ragni *et al.*, "Data augmentation for low resource languages," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 810–814.

[21] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2613–2617.

[22] Q. Kong *et al.*, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[23] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–15.

[24] B. Schuller *et al.*, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.