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

Tomoya Koike\textsuperscript{1}, Student Member, IEEE, Kun Qian\textsuperscript{1,2}, Senior Member, IEEE, Björn W. Schuller\textsuperscript{3,4}, Fellow, IEEE, and Yoshiharu Yamamoto\textsuperscript{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 highlight 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 (\textit{p} < .05 by one-tailed \textit{t}-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 mimicking 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
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 \(2\pi\). 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.

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/

<table>
<thead>
<tr>
<th>Database</th>
<th>Recordings</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Duration [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT</td>
<td>409</td>
<td>117</td>
<td>292</td>
<td>32.6</td>
</tr>
<tr>
<td>AAD</td>
<td>490</td>
<td>386</td>
<td>104</td>
<td>7.9</td>
</tr>
<tr>
<td>AUTH</td>
<td>7</td>
<td>7</td>
<td>24</td>
<td>49.4</td>
</tr>
<tr>
<td>UHA</td>
<td>55</td>
<td>27</td>
<td>28</td>
<td>15.1</td>
</tr>
<tr>
<td>DLUT</td>
<td>2,141</td>
<td>1,958</td>
<td>183</td>
<td>23.1</td>
</tr>
<tr>
<td>SUA</td>
<td>114</td>
<td>80</td>
<td>34</td>
<td>33.1</td>
</tr>
</tbody>
</table>
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 achieve 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
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**