Audio for Audio is Better? An Investigation on Transfer Learning Models for Heart Sound Classification

Tomoya Koike[1], *Student Member, IEEE*, Kun Qian[1], *Member, IEEE*,
Qiuqiang Kong[2], *Student Member, IEEE*, Mark D. Plumbley[2], *Fellow, IEEE*
Björn W. Schuller[3,4], *Fellow, IEEE*, Yoshiharu Yamamoto[1], *Member, IEEE*

***Abstract*— Cardiovascular disease is one of the leading factors for death cause of human beings. In the past decade, heart sound classification has been increasingly studied for its feasibility to develop a non-invasive approach to monitor a subject's health status. Particularly, relevant studies have benefited from the fast development of wearable devices and machine learning techniques. Nevertheless, finding and designing efficient acoustic properties from heart sounds is an expensive and time-consuming task. It is known that transfer learning methods can help extract higher representations automatically from the heart sounds without any human domain knowledge. However, most existing studies are based on models pre-trained on images, which may not fully represent the characteristics inherited from audio. To this end, we propose a novel transfer learning model pre-trained on large scale audio data for a heart sound classification task. In this study, the PhysioNet CinC Challenge Dataset is used for evaluation. Experimental results demonstrate that, our proposed pre-trained audio models can outperform other popular models pre-trained by images by achieving the highest unweighted average recall at 89.7 %.**

I. INTRODUCTION

Auscultation using a stethoscope is an efficient, inexpensive, and convenient way for making an early diagnosis of cardiovascular disease (CVD), which accounts for approximately 45 % of all deaths annually in Europe [1]. Nevertheless, it is time-consuming and inefficient to train a sufficient number of physicians to be qualified in using a stethoscope for clinical practice [2]. The components of a heart sound include the first (S1) and the second sound (S2) as normal sounds, while the third and the fourth heart sounds, i. e., S3 and S4, often correspond to murmurs, and ejection clicks, usually refer to some disease, or anomaly [3].

Within the fast development in signal processing and machine learning, heart sound classification has been increasingly studied in the past decades [3]. In recent feature extraction approaches, wavelet analysis, time-frequency analysis using short time fourlier transform (STFT), unsupervised learning and other methods are often found [4]. As a classifier, hidden Markov model (HMM), k-nearest neighbour, support vector machines, random forest, multi-layer perceptron, deep neural network, convolutional neural network, recurrent neural network, and other classifiers have been used in previous research [4].

Existing studies have shown encouraging results that may lead into a promising future direction on developing non-invasive methods for automatically monitoring the heart status. On the other hand, finding and designing efficient acoustic features for heart sounds needs expensive domain knowledge of human experts. Moreover, to make the current machine learning-based approaches feasible in clinical practice, a large number of expert annotations are needed, which is another difficult issue for almost all biomedical engineering fields. Motivated by the success of transfer learning (TL) in computer vision [5], natural language processing [6], and speech recognition [7], TL-based methods are now proving another paradigm for extracting higher representations from heart sound without any human expert domain knowledge [8]. Nonetheless, most existing TL-based models are pre-trained on images, such as ImageNet [9] rather than on audio data. To explore the TL capacity of a most recently released deep model pre-trained on large scale audio data, such as the Large-Scale Pretrained Audio Neural Networks (PANNs) for audio pattern recognition [10], we introduce PANNs for the heart sound classification task. Our hypothesis is that the deep model pre-trained on audio may catch more inherited characteristics from heart sounds than models pre-trained on images.

This research work was partially supported by 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. 17H00878) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P. R. China. Tomoya Koike and Kun Qian are the *Corresponding Authors*.

[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]Qiuqiang Kong and Mark D. Plumbley are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK. {q.kong, m.plumbley}@surrey.ac.uk

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

The main contributions of this work can be summarised as: First, we introduce a novel deep learning model pre-trained on large scale audio data into the paradigm of TL for heart sound classification. To the best of our knowledge, it is the first time an audio based pre-trained TL model is used for heart sound classification. Second, we investigate and compare the proposed models with other state-of-the-art TL models on their capacity to extract higher representations from heart sound. Third, we compare two popular inputs, the spectrogram and the Log Mel spectrogram, for their performance on the interpretation of the heart status. This paper is organised as follows: The database used and the method will be introduced in Section II. Experimental results and discussion will be given in Section III and Section IV, respectively. Finally, we conclude this work in Section V.

II. Materials and Methods

A. Dataset

In this study, the open-access heart sound database, PhysioNet CinC Challenge database [11] is used. This database was collected from nine medical centres, which includes 2 435 Phonocardiogram (PCG) recordings from healthy subjects and 1 297 PCG recordings from the patients suffering from a variety of heart valve diseases and coronary artery diseases. The length of the PCG recordings is ranging from several seconds to minutes. Public data for training and private data for scoring were used in the CinC Challenge. We used the former one, including 3 240 PCG instances. Minimum, maximum, and mean length of those PCG are 5.3, 121.9 and 22.4 seconds, respectively. All instances were resampled with 2 000 Hz.

B. Transfer Learning from Images

Convolutional neural networks (CNNs) have given good results in the computer vision field in recent years, but they need a high amount and variety of data to train with high computing costs. On this background, TL is commonly used with CNNs pre-trained on large datasets like ImageNet. There are several available pre-trained models for image tasks: VGG [12], MobileNet [13], ResNet [14], and ResNeXt [15]. In the ImageNet 2014 challenge, the model known as VGG reached the second place in the classification track with 3×3 convolution filters [12]. This model is deeper than the previous models at that time with 16 ˜ 19 convolutional layers. Residual blocks are proposed on the background that deeper CNNs achieve higher performance while it is challenging to train them due to vanishing gradient [14]. Residual blocks are composed of two pathways: via the convolutional layer, and via direct input as a shortcut. The shortcut path takes no extra parameters and matrix calculation, making backpropagation easier at the same time. In the ResNets, convolutions with 3×3 filters cost heavily on computing. To reduce the computational cost, MobileNet V1 factorised a standard convolution into two convolutions: Depthwise convolution and pointwise convolution by 1×1 filters [13]. Inverted residual blocks, known as bottleneck layers, were added in MobileNet V2 [16]. MobileNet V2 has fewer parameters than MobileNet V1. To obtain stronger representational power than ResNet, ResNeXt [15] added group convolutions and reduced channel-wise compression rate, keeping the number of parameters on par with ResNet. Weakly supervised learning with ResNeXt was proposed in [17]. It is pre-trained in a weakly-supervised fashion on 940 million public images with 1.5 k hashtags matching with 1 000 ImageNet1K synsets, followed by fine-tuning on ImageNet1K dataset.

TABLE I: Topology of the PANN CNN14 model.

Log Mel spectrogram 120000 frames × 64 mel bins
$(3\times3@64, BN, ReLU)\times2$, Pooling 2×2
$(3\times3@128, BN, ReLU)\times2$, Pooling 2×2
$(3\times3@256, BN, ReLU)\times2$, Pooling 2×2
$(3\times3@512, BN, ReLU)\times2$, Pooling 2×2
$(3\times3@1024, BN, ReLU)\times2$, Pooling 2×2
$(3\times3@2048, BN, ReLU)\times2$, Global pooling
FC 2048, ReLU
FC 2, Softmax

C. Transfer Learning from Audio: PANNs

To provide a generalised model in the audio pattern recognition field, large-scale pre-trained audio neural networks (PANNs) were proposed [10]. A wide range of convolutional neural networks were pre-trained on to classify 527 sound classes. Particularly, a 14-layer CNN was transferred and fine-tuned on several audio pattern tasks. Their CNN pre-trained on AudioSet is generalised well in many audio pattern recognition tasks. We used CNN14 from [10] which has five blocks of 3 x 3 convolutional filters, batch normalisation and ReLU as shown in Table I. The number of frames and output size were modified to fit the CinC dataset. The number after symbol @ indicates the number of feature maps. "BN" and "FC" indicate batch normalisation and fully connected layer, respectively. The whole system structure is shown in Fig. 1. The loss function with which all of those CNN models are fine-tuned is binary cross-entropy or log loss which is defined as:

$$Log\,Loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)], \quad (1)$$

where n is the number of instances, $\hat{y}_i$ is the predicted probability of abnormal label, y_i is 1 if label is abnormal and 0 if label is normal.

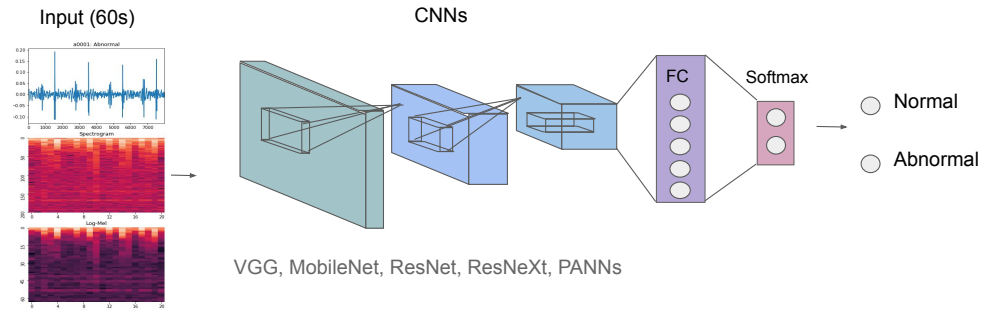


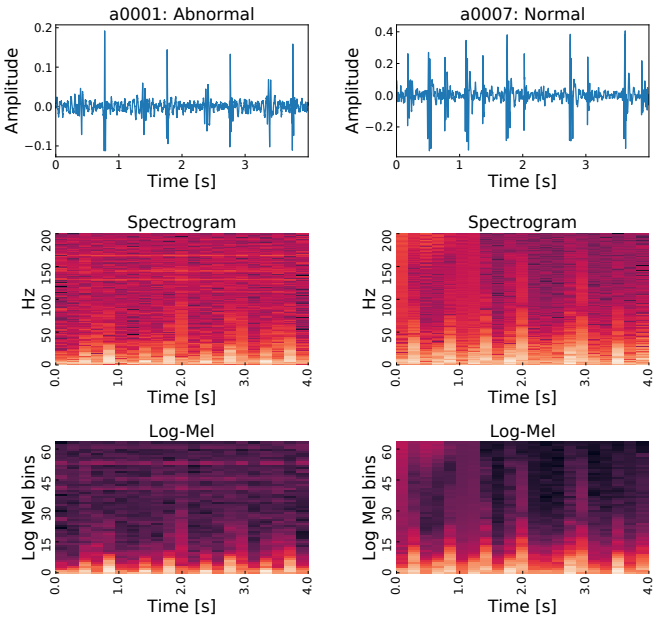
Fig. 1: Diagram of the TL-based method for heart sound classification. The original heart sound audio data is transformed to Log Mel or 'normal' spectrograms as the inputs at first. Then, the pre-trained deep models containing multiple convolutional layers will be used to extract higher representations from the inputs. Finally, the prediction will be made by a fully connected (FC) layer and a softmax layer for targeting the inputs to the classes of heart sound, i. e., 'normal' or 'abnormal'.

III. Experimental Results

A. Pre-processing

There were two pre-processing methods to obtain feature maps from raw waveform inputs used in the CinC Challenge 2016: spectrogram and Log Mel spectrogram. The input wave and the pre-processed feature maps are shown in Fig. 2. In order to get frequency features without losing temporal change, the spectrogram was obtained by splitting waveform data into a window length, choosing the first split waveform, calculating short-time Fourier transform (STFT) and executing accordingly to the end of the split waveform. This pre-processing was used for example in the CinC Challenge 2016 [18]. After reaching a spectrogram representation by the process above, the Log Mel spectrogram is obtained by multiplying the Mel filter bank with the spectrogram and applying the logarithm. The Mel filter bank increases the size of the passing frequency range as the frequency increases. Hence, features in higher frequency range in the spectrogram are relatively coarse as compared to those in lower frequency with respect to spectral resolution. This pre-processing was used for example in the CinC Challenge in [19]. The raw heart sound waveform, spectrogram, and Log Mel spectrogram are shown in Fig. 2.

Fig. 2: Waveform (top), spectrogram (middle), and Log Mel spectrogram (bottom) of the heart sound audio samples (in 4 s) annotated as 'abnormal' (left), and 'normal' (right).



B. Experimental Setup

The length of the CinC data ranges from several seconds to several minutes. To make it the same length, we cut the start and end of records when their length is longer than 60 seconds, and if the length is shorter than 60, we padded it at the centre. We split them into 3:1:1 as train, development, and test datasets, keeping in each dataset the same normal/abnormal ratio with the original data at about 4:1. In order to train CNNs successfully, each label was sampled equally. Hyper parameters were kept the same among all CNNs: learning rate = 0.0001, STFT window size = 400, STFT window stride = 400, and the number of Mel bins = 64 under 1000 Hz. Considering the imbalanced data distribution of the PhysioNet CinC Challenge database, we use specificity, sensitivity, F1 score and the unweighted average recall (UAR) [20] as the evaluation metrics. UAR is defined as:

$$\mathrm{UAR} = \frac{\sum_{i=1}^{N_c} \mathit{Recall}_i}{N_c}, \tag{2}$$

where N_c is the number of classes ($N_c = 2$ in this study).

C. Results

The experimental results are shown in Table II and Table III. CNNs except PANNs accept spectrogram and Log Mel spectrogram as an input, while PANNs accept a raw waveform. PANNs include the STFT layer and the Mel filter bank layer in their weights pre-trained on AudioSet. Therefore, we cannot remove the Mel filter