Blind Monaural Source Separation on Heart and Lung Sounds Based on Periodic-Coded Deep Autoencoder

Kun-Hsi Tsai[1], Wei-Chien Wang[1], Chui-Hsuan Cheng[1], Chan-Yen Tsai[1], Jou-Kou Wang[2], Tzu-Hao Lin[3], Shih-Hau Fang[4], Li-Chin Chen[5], Yu Tsao[5]

Abstract—**Auscultation is the most efficient way to diagnose cardiovascular and respiratory diseases. To reach accurate diagnoses, a device must be able to recognize heart and lung sounds from various clinical situations. However, the recorded chest sounds are mixed by heart and lung sounds. Thus, effectively separating these two sounds is critical in the pre-processing stage. Recent advances in machine learning have progressed on monaural source separations, but most of the well-known techniques require paired mixed sounds and individual pure sounds for model training. As the preparation of pure heart and lung sounds is difficult, special designs must be considered to derive effective heart and lung sound separation techniques. In this study, we proposed a novel periodicity-coded deep auto-encoder (PC-DAE) approach to separate mixed heart-lung sounds in an unsupervised manner via the assumption of different periodicities between heart rate and respiration rate. The PC-DAE benefits from deep-learning-based models by extracting representative features and considers the periodicity of heart and lung sounds to carry out the separation. We evaluated PC-DAE on two datasets. The first one includes sounds from the Student Auscultation Manikin (SAM), and the second is prepared by recording chest sounds in real-world conditions. Experimental results indicate that PC-DAE outperforms several well-known separation works in terms of standardized evaluation metrics. Moreover, waveforms and spectrograms demonstrate the effectiveness of PC-DAE compared to existing approaches. It is also confirmed that by using the proposed PC-DAE as a pre-processing stage, the heart sound recognition accuracies can be notably boosted. The experimental results confirmed the effectiveness of PC-DAE and its potential to be used in clinical applications.**

Index Terms—**Blind Monaural Source Separation, Deep Autoencoder, Deep Neural Networks, Heart Sound, Lung Sound, Phonocardiogram, Periodic Analysis.**

Manuscript submitted December 31th, 2019.

K.-H. Tsai, W.-C. Wang, C.-H. Cheng, C.-Y. Tsai are with the Medical Department, Imediplus Inc. Taipei, Taiwan ({peter.tsai; ethan.wang;hammer.cheng;mike.tsai}@imediplus.com).
L.-C. Chen and Y. Tsao are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan. ({li.chin; yu.tsao)@citi.sinica.edu.tw).
J.-K. Wang is with Department of Pediatrics, National Taiwan University Hospital, Taipei, Taiwan (jkww@ntuh.gov.tw).
S.-H. Fang is with Department of Electric Engineering, Yuan Ze University, Taoyuan, Taiwan (shfang@saturn.yzu.edu.tw)
T.-H. Lin is with Biodiversity Research Center, Academia Sinica, Taiwan (schonkopf@gmail.com)

I. Introduction

Recently, biological acoustic signals have been enabling various intelligent medical applications. For example, the biological acoustic signals of the heart and lung can facilitate tasks such as diagnosing the cardiovascular and respiratory diseases, and monitoring the sleep apnea syndrome [1-8]. Previous studies have already investigated the physical models of the heart and lung sound generation and classification mechanisms. For example, signal processing approaches (e.g., normalized average Shannon energy [9] and high-frequency-based methods [10]) and machine-learning-based models (e.g., neural network (NN) classifiers [11] and decision trees [12]) have been used to perform heart disease classification based on acoustic signals. In addition, the information of S1–S2 and S2–S1 intervals has been adopted to further improve the classification accuracies [12], [13]. On the other hand, Gaussian mixture model [13] NN classifiers [14], and support vector machines[15] along with various types of acoustic features (e.g., power spectral density values, Hilbert-Huang transform[16]) have been utilized to carry out lung sound recognition [17, 18]. However, medical applications using such biological acoustic signals still face several challenges.

To reach accurate recognition, sound separation is one of the most important pre-processing. Because the measured signal is usually a mixed version of the heart and lung sounds, and pure heart/lung acoustic signals is generally not accessible, effectively separating heart and lung sounds is very challenging. The frequency ranges of normal heart sounds (first(S1) and second(S2) heart sound) is mainly 20-150Hz, and some high-frequency murmurs may reach to 100-600Hz, or even to 1000Hz [19]. On the other hand, the frequency range of normal lung sounds is 100-1000Hz (tracheal sounds range from 850Hz to 1000Hz), abnormal lung sound as adventitious sounds of wheeze span a wide range of frequencies variation of 400-1600Hz, and the range for crackle and rales is 100-500Hz [20, 21]. Therefore, the frequency range of the heart and lung sounds can be highly overlapped. This results in interference between

Samples of output audios are available at: https://weichian0920.github.io/

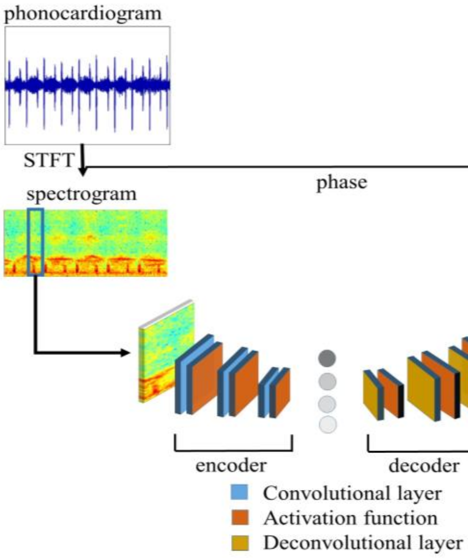


Fig. 1. The convolutional deep autoencoder (DAE(C)) architecture.

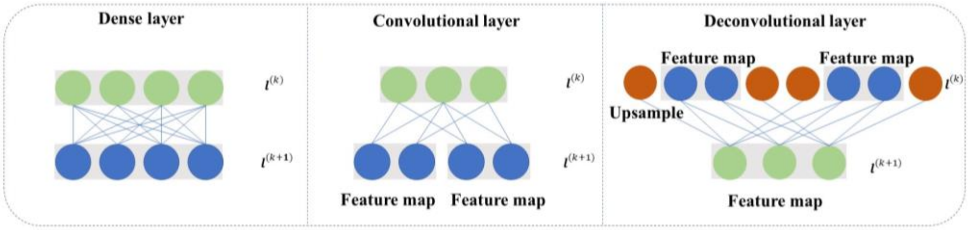


Fig. 2. Relation between hidden layers in a fully connected layer, convolutional layer, and deconvolutional layer.

the acoustic signals and may degrade the auscultation and monitoring performance. With an increasing demand for various acoustic-signal-based medical applications, effective heart and lung sound separation techniques have become fundamental, although challenging.

Sound separation techniques for heart and lung have been studied extensively, and numerous methods have been proposed so far. For example, the study [22-26] focuses on the adaptive filtering approach while Mondal et al. [27, 28] use the empirical mode decomposition methods. Hossain and Hadjileontiadis et al. [29, 30] proposed to use the discrete wavelets transform approach to filter interference. Pourazad et al. [31] derived an algorithm that transforms the signal to time-frequency domain (STFT) and combined with the continuous wavelets transform (CWT) to filter out heart sound components by a band-pass filter.

However, the above-mentioned traditional filtering approaches encounter difficulties due to the overlapped frequency bands. The works in [32-34] proposed the blind source separation algorithms, including independent component analysis (ICA) and its extensions, in which the prior knowledge of sources is not required. Nevertheless, the ICA-based methods require at least two sensors and thus, do not work for the devices having only single-channel [35-37]. The assumption of independence between heart sound sources is somehow optimistic.

Recently, the supervised monaural (single-channel) nonnegative matrix factorization (NMF) was adopted to separate different sources [35, 38]. It was recognized for its capability of handling overlapping frequency bands [39, 40]. More recently, deep learning approaches have been used for source separation [40-43]. Although these deep models directly dismantle the mixture source into the target ones and outperform the NMF approach, those frameworks were subject to supervised training data. However, in biomedical applications, the training data of pure heart/lung acoustic signals is difficult or too expensive to measure.

To overcome the mentioned challenges, this paper proposes a periodicity-coded deep autoencoder (PC-DAE) approach, an unsupervised-learning-based mechanism to effectively separate the sounds of heart and lung sources. The proposed algorithm first adopts the DAE model [40, 44-46] to extract highly expressive representations of the mixed sounds. Next, by applying the modulation frequency analysis (MFA) [47] on the latent representations, we can group the neurons based on their properties in the modulation domain and then perform separation on the mixed sound. The advantage of PC-DAE is that the labeled training data (more specifically, paired mixed sounds and individual pure sounds) are not required as compared to the typical learning-based approaches. It benefits from the periodicity structure to provide superior separation performance than the traditional methods.

The remainder of this paper is organized as follows. In Section 2, we will review the NMF and DAE algorithms. In Section 3, the proposed PC-DAE will be introduced in detail.

In Section 4, we present the experimental setup and results, where two datasets were designed and used to test the proposed PC-DAE model. The first one is phonocardiogram signals from the Student Auscultation Manikin (SAM) database) [48] , and the second one is prepared in a real-world condition. Experimental results confirm the effectiveness of PC-DAE to separate the mixed heart-lung sounds with outperforming related works, including direct-clustering NMF (DC-NMF) [35], PC-NMF [49], and deep clustering (DC) [45], in terms of three standardized evaluation metrics, qualitative comparisons based on separated waveforms and spectrograms, and heart sound recognition accuracy.

II. Related Works

Numerous methods have been proposed to separate the heart and lung sound signals. Among them, the NMF is a notable one that has been applied to separate different sounds [35, 38]. The DAE model is another well-known approach. Based on the model architecture, the DAE can be constructed by a fully connected architecture, termed DAE(F), or by a fully convolutional architecture, termed DAE(C). In this section, we provide a review of the NMF algorithm, DAE(F), and DAE(C) models.

A. Non-negative matrix factorization (NMF)

The conventional NMF algorithm factorizes the matrix $\boldsymbol{V}$ into two matrices, a dictionary matrix $\boldsymbol{W}$ and an encoded matrix $\boldsymbol{H}$. The product of the $\boldsymbol{W}$ and $\boldsymbol{H}$ approximates matrix $\boldsymbol{V}$. All the matrices entries are nonnegative. The NMF-based source separation can be divided into two categories, namely supervised (where individual source sounds are provided) and unsupervised (where individual source sounds are not accessible). For supervised NMF-based approaches, a pre-trained, fixed spectral matrix $\boldsymbol{W}^S$, where $\boldsymbol{W}^S = [\boldsymbol{W}_1^S \dots \boldsymbol{W}_M^S]$, and M is the number of sources, which consists of the characters of each sound source is previously required [35, 50]. To process NMF, first, the recording that consists of multiple sounds was factorized by NMF into $\boldsymbol{W}^S$ and $\boldsymbol{H}^T$. Then $\boldsymbol{H}^T$ is divided into M blocks: $\boldsymbol{H}^T = [\boldsymbol{H}_1^T \dots \boldsymbol{H}_M^T]$, where M is the number of sources. Through multiplying $\boldsymbol{W}_i^S$ and $\boldsymbol{H}_i^T$ (i=1,...M), we obtain individual sound sources.

For unsupervised NMF-based approaches, since individual source sounds are not available, some statistical assumptions must apply. An intuitive approach is to cluster the vectors in $\boldsymbol{H}$ to several distinct groups. A particular sound can be reconstructed by a group of vectors in $\boldsymbol{H}$ along with $\boldsymbol{W}$. The work of Lin et al [49], on the other hand, designed PC-NMF using another concept, which is to incorporate the periodicity property of distinct source sounds into the separation framework. More specifically, PC-NMF considers the encoded matrix $\boldsymbol{H}^T$ as the time vector and uses the nature of periodical differences to separate the biological sounds. Because heart sound and lung sounds are different in periodic characters (heart rate and respiration rate are very different), the mixed heart-lung sound is separated through a PC-NMF model, as will be presented in Section 4.

B. Deep Autoencoder (DAE)

The DAE has two components, an encoder $E(\cdot)$ and a decoder $D(\cdot)$. Figure 1 shows the architecture of a DAE(C) model. Consider the encoder and decoder to have K_E and K_D layers, respectively, the total number of layers in the DAE is $K_{All} = K_E + K_D$. The encoder encodes the input $\boldsymbol{x}$ to the middle latent space $\boldsymbol{l}^{(K_E)}$ ($\boldsymbol{l}^{(K_E)} = E(\boldsymbol{x})$), and the decoder reconstructs the input by ($\boldsymbol{y} = D(\boldsymbol{l}^{(K_E)})$). The reconstructed output $\boldsymbol{y}$ is expected to be approximately equal to $\boldsymbol{x}$. The mean squared errors (MSE) is generally used to measure the difference between $\boldsymbol{y}$ and $\boldsymbol{x}$. Minimizing the value of MSE is the goal to train the DAE model. As mentioned earlier, by using fully connected and fully convolutional architectures, we can build DAE(F) and DAE(C), respectively [51-53]. Fig. 2 shows the neuron connections of the *k-th* and (k+1)-th layers for the two types of DAE. Fig. 2(a) presents the fully-connected layer, where each neuron in the (k+1)-th layer is fully-connected with all neurons in the k-th layer. Fig. 2 (b) and (c), respectively, present the convolutional and deconvolutional connections, where each neuron in the (k+1)-th layer is partially-connected with the neurons in the k-th layer. As can be seen from Fig. 2(a), the DAE(F) form the encoder and decoder by fully-connected units, which is shown in Eqs. (1) and (2), $\boldsymbol{W}_E^{(k)}$ and $\boldsymbol{W}_D^{(k)}$ represent the encoding and decoding matrix, $\boldsymbol{b}_E^{(k)}$ and $\boldsymbol{b}_D^{(k)}$ are the vectors of biases:

$$\begin{aligned} \boldsymbol{l}^{(1)} &= \sigma\left(\boldsymbol{W}_E^{(0)}\boldsymbol{x} + \boldsymbol{b}_E^{(0)}\right) \\ \boldsymbol{l}^{(k+1)} &= \sigma\left(\boldsymbol{W}_E^{(k)}\boldsymbol{l}^{(k)} + \boldsymbol{b}_E^{(k)}\right) \quad k = 1, \dots, K_E \text{ -1}, \end{aligned} \tag{1}$$

where $\boldsymbol{l}^{(k)} \in R^{M\times 1}$, and M stands for the total number of neurons in the latent space. For the decoder, we have

$$\begin{aligned} \boldsymbol{l}^{(k+1)} &= \boldsymbol{\sigma}\left(\boldsymbol{W}_D^{(k)}\boldsymbol{l}^{(k)} + \boldsymbol{b}_D^{(k)}\right), \ k = K_D \dots, (K_{All} \text{ -1}) \\ \boldsymbol{x} &= \boldsymbol{\sigma}\left(\boldsymbol{W}_D^{(K_{All})}\boldsymbol{l}^{(K_{All})} + \boldsymbol{b}_D^{(K_{All})}\right). \end{aligned} \tag{2}$$

In DAE(C), the encoder is formed by convolutional units, as shown in Eq. (3), that executes the convolutional function $F_{Conv}(\cdot)$. Each encoded layer has J filters: $\{\boldsymbol{W}_1, \dots, \boldsymbol{W}_J\}$; $\boldsymbol{W}_j \in R^{L\times 1}$, L is the kernel size, and $\boldsymbol{W}_{ji}$ is the i-th channel of $\boldsymbol{W}_j$, where $\boldsymbol{W}_{ji} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_I)$. Each neuron in the ($k$+$1$)-th layer's feature map, $\boldsymbol{l}_j^{(k+1)}$, is the summation of the element-wised product of $\boldsymbol{W}_j$ and receptive field of all previous feature maps $\boldsymbol{l}^{(k)}$ by convolution operation, and $\boldsymbol{b}_j^{(k)}$ denotes the bias terms. The corresponding convolution operation is shown in Fig. 3 (a). The decoder is formed by a deconvolutional unit, as shown in Eq. (4). During deconvolution, all of the k-th layer's feature maps $\boldsymbol{l}^{(k)}$ first go through the zero-padding and then deconvolution processes (with function $F_{Deconv}(.)$). Each decoded layer has J filters: $\{\boldsymbol{W}_1, \dots, \boldsymbol{W}_J\}$; $\boldsymbol{W}_j \in R^{L\times 1}$, L is the kernel size, and $\boldsymbol{W}_{ji}$ is the i-th channel of $\boldsymbol{W}_j$, where $\boldsymbol{W}_{ji} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_I)$. Each neuron in the ($k$+$1$)-th layer, $\boldsymbol{l}_j^{(k+1)}$, is the summation of the element-wised product of $\boldsymbol{W}_j$ and receptive field of all previous feature maps $\boldsymbol{l}^{(k)}$ by deconvolution operation, and $\boldsymbol{b}_j^{(k)}$ denotes the bias terms. The corresponding deconvolution operation is shown in Fig. 3 (b).

$$\begin{aligned} \boldsymbol{l}_j^{(1)} &= \sigma\left(F_{Conv}(\boldsymbol{W}_{ji}^{(0)}, \boldsymbol{x}) + \boldsymbol{b}_{E_j}^{(0)}\right) \\ \boldsymbol{l}_j^{(k+1)} &= \sigma\left(\sum_{i=1}^{I} F_{Conv}(\boldsymbol{W}_{ji}^{(k)}, \boldsymbol{l}^{(k)}) + \boldsymbol{b}_{E_j}^{(k)}\right) \end{aligned} \qquad (3)$$

where $\boldsymbol{l}_j^{(k)} \in R^{M \times N}$ is the j-th feature map in the k-th layer, and I is the total number of channels. For the decoder, we have

$$\begin{aligned} \boldsymbol{l}_j^{(k+1)} &= \sigma\left(\sum_{i=1}^{I} F_{Deconv}(\boldsymbol{W}_{ji}^{(k)}, \boldsymbol{l}^{(k)}) + \boldsymbol{b}_j^{(k)}\right) \\ \boldsymbol{x} &= \sigma\left(F_{Deconv}(\boldsymbol{W}_{ji}^{(K_{All})}, \boldsymbol{l}^{(K_{All})}) + \boldsymbol{b}_j^{(K_{All})}\right) \end{aligned} \qquad (4)$$

where K_{All} denotes the total number of layers in the DAE(C).

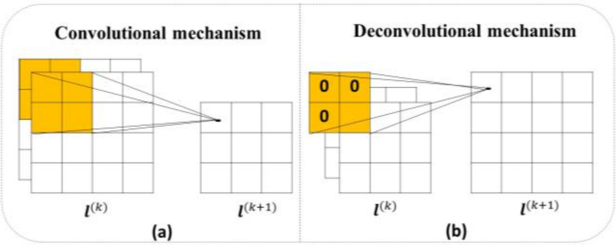


Fig. 3 (a) Convolutional and (b) deconvolutional operations.

III. The Proposed Method

The proposed PC-DAE is a DAE-based unsupervised sound source separation method. When performing separation, the recorded sounds are first transformed to spectral-domain and phase parts via short-time Fourier transform (STFT). The spectral features are converted to log power spectrum (LPS) [52], where $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \dots, \boldsymbol{x}_N]$ denotes the input, and $\boldsymbol{N}$ is the number of frames of $\boldsymbol{X}$. Then the DAE encodes the mixed heart-lung LPS by $E(\cdot)$ to convert $\boldsymbol{X}$ to the matrix of latent representations, $\boldsymbol{L}^{(K_E)} = [\boldsymbol{l}_1^{(K_E)}, \dots, \boldsymbol{l}_n^{(K_E)}, \dots, \boldsymbol{l}_N^{(K_E)}]$. The decoder, $D(\cdot)$, then reconstructs the latent representations back to original spectral features. The back-propagation algorithm [54] is adopted to train the DAE parameters to minimize the MSE scores. Because the input and output are the same, the DAE can be trained in an unsupervised manner.

With the trained DAE, the periodic analysis is applied to the latent representations to identify two disjoint portions of neurons corresponding to heart and lung sounds. The basic concept is to consider the temporal information of different periodic sources. Moreover, to classify the temporal information by periodicity, the coded matrix is transformed into periodic coded matrix $\boldsymbol{P}$ via modulation frequency analyzer (MFA). Here, we adopted the discrete Fourier transform (DFT) to perform MFA. The periodic coded matrix presents clear periodicity characteristics. Owing to heart sound and lung sound have different periodicity, the coded matrix can be separated to heart coded matrix and lung coded matrix from the whole encoded matrix, $\boldsymbol{P}$. Afterwards, each source coded matrix is transformed by the decoder and reconstructed to obtain the LPS sequences of the separated heart sound $\boldsymbol{Y}^{heart}$ and lung sound $\boldsymbol{Y}^{lung}$. The output LPS features are then converted back to waveform-domain signals by applying inverse short-time Fourier transform (ISTFT).

A. Periodic Analysis Algorithm

In this section, we present the details of the MFA. Fig. 4 illustrates the overall PC-DAE framework. First, we train a DAE(F) or DAE(C) model with the encoder and decoder as shown in Eqs. (1) and (2) or Eqs. (3) and (4), respectively. Then, we input the sequence of mixed heart-lung sounds, $\boldsymbol{X}$, to obtain the latent representations. The collection of latent representations and the time sequence are the matrix $\boldsymbol{L}$={$\boldsymbol{l}_1^{(K_E)}, \boldsymbol{l}_2^{(K_E)}, \dots \boldsymbol{l}_N^{(K_E)}$}. Thus, we obtain

$$\begin{aligned} \boldsymbol{L} &= \left[E(\boldsymbol{x}_1) \quad \dots \quad E(\boldsymbol{x}_n) \quad \dots \quad E(\boldsymbol{x}_N)\right] \\ &= \left[\begin{bmatrix} l_{11}^{(K_E)} \\ \vdots \\ l_{j1}^{(K_E)} \\ \vdots \\ l_{M1}^{(K_E)} \end{bmatrix} \dots \begin{bmatrix} l_{1n}^{(K_E)} \\ \vdots \\ l_{jn}^{(K_E)} \\ \vdots \\ l_{Mn}^{(K_E)} \end{bmatrix} \dots \begin{bmatrix} l_{1N}^{(K_E)} \\ \vdots \\ l_{jN}^{(K_E)} \\ \vdots \\ l_{MN}^{(K_E)} \end{bmatrix}\right], \end{aligned} \qquad (5)$$

where $\boldsymbol{L} \in R^{M \times N}$, j is the neuron index, where $1 \leq j \leq M$, and n is the time stamp, where $1 \leq n \leq N$, and N is the total number of time stamps

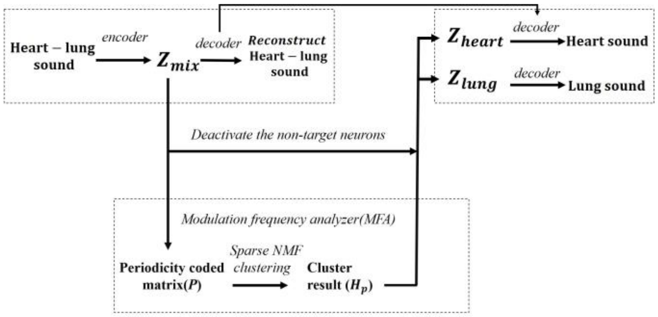


Fig. 4. The PC-DAE Framework.

We assume that among the latent representations, some neurons are activated by heart sound and the others activated by lung sounds. Based on this assumption, we can separate mixed heart-lung sounds in the latent representation space. To determine whether each neuron is activated either by heart or lung sound, we transpose the original $\boldsymbol{L}$ to obtain $\boldsymbol{Z}^{mix} = \boldsymbol{L}^T$ (T denotes matrix transpose). Thus, we obtain

$$\begin{aligned} \boldsymbol{Z}^{mix} &= [\boldsymbol{z}_1^{mix}, \dots, \boldsymbol{z}_j^{mix}, \dots, \boldsymbol{z}_M^{mix}], \\ \text{where} \quad & \\ \boldsymbol{z}_j^{mix} &= [l_{11}^{(K_E)}, \dots, l_{jn}^{(K_E)}, \dots, l_{jN}^{(K_E)}]^T \end{aligned} \qquad (6)$$

With $\boldsymbol{Z}^{mix}$, we intend to cluster the entire set of neurons into two groups, one group corresponding to heart sound and the other to lung sounds. More specifically, when pure heart sound is inputted to the DAE, only one group of neurons corresponding to the heart sounds is activated, and the other group corresponding to the lung sounds is deactivated. When the pure lung sound is inputted to the DAE, on the other hand, the group of neurons corresponding to the lung sounds is activated, and the other group corresponding to the heart sounds is deactivated. The strategy to determine these two groups of neurons is based on the periodicity of heart and lung sounds.

Algorithm 1 shows the detailed procedure of periodic analysis. To analyze the periodicity of each submatrix $\boldsymbol{z}_j^{mix}$, we form the periodic coded matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_j, \dots, \boldsymbol{p}_M]$ by applying the MFA on $\boldsymbol{z}_j^{mix}$, as shown in Eq. (7).

$$\boldsymbol{p}_j = |\text{MFA}(\boldsymbol{z}_j^{mix})|. \qquad (7)$$

When we used DFT to carry out MFA, we have $\boldsymbol{p}_j \in R^{(N/2+1)}$, and $\boldsymbol{P}$ can be clustered into two groups. There are numerous clustering approaches available, and we used the sparse NMF clustering method to cluster the vectors in $\boldsymbol{P}$ into two groups [55]. Eq. (8) shows the clustering process by NMF, which is also achieved by minimizing the error function. On the basis of the largest score in the encoding matrix, $\boldsymbol{H}_p$, of the transposed $\boldsymbol{P}^T$**,** the clustering assignment of $\boldsymbol{Z}^{mix}$ can be determined.

$$\boldsymbol{H}_p = \arg min[\| \boldsymbol{P}^T - \boldsymbol{W}_p \boldsymbol{H}_p \|^2 + \lambda \| \boldsymbol{H}_p \|], \qquad (8)$$

where $\boldsymbol{W}_p$ represents the cluster centroids, $\boldsymbol{H}_p = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_j, \dots, \boldsymbol{h}_M]$ represents the cluster membership, $\boldsymbol{h}_j \in \boldsymbol{R}^{k \times 1}$, k is set as the cluster amount of the basis, λ represents the sparsity penalty factor, $|| \cdot ||$ represents the L1-norm, and $\|\cdot\|_F^2$ represents the Frobenius distance.

Algorithm 1 : MFA on coded matrix

Input: mixed heart-lung coded matrix $\boldsymbol{Z}^{mix}$, where $\boldsymbol{Z}^{mix} \in R^{M \times N}$

Output: heart coded matrix $\boldsymbol{Z}^{heart}$, lung coded matrix $\boldsymbol{Z}^{lung}$

```
for $j$ = 1 to $M$ do
    $\boldsymbol{p}_j = |$ MFA($\boldsymbol{z}_j^{mix}$)|
end for

Perform clustering on vectors $[\boldsymbol{p}_1, \dots, \boldsymbol{p}_M]$ in $\boldsymbol{P}$
Obtain labels of $\boldsymbol{P}$: $\boldsymbol{c} = [c_1, \dots, c_M]$, where there are only
  two labels of $c_j$ {heart or lung}.
Assign the label $c_j$ to $\boldsymbol{z}_j^{mix}$

Set $\boldsymbol{\sigma}_{min} \in R^{M \times N}$, where $\boldsymbol{\sigma}_{min}$ is a vector whose coeffi-
  cients are the latent neuron's minimum values
foreach $t$ = [heart; lung] do
    Initialize $\boldsymbol{Z}^t$ = $\boldsymbol{Z}^{mix}$
    for $j$ (1 to $M$) do
        if $c_j \neq t$ then
            do $\boldsymbol{z}_j^t$ = $\boldsymbol{\sigma}_{min}$
        end if
    end for
    return $\boldsymbol{Z}^t$
end foreach
```

On the basis of the $\boldsymbol{h}_j$ of encoding matrix $\boldsymbol{H}_p$, the clustering results $C = [c_1, \dots, c_j, \dots, c_M]$ is determined by the largest score of h_j . In this case, $c_j \in$ {heart, lung}, and the cluster results assign to $\boldsymbol{z}_j^{mix}$. According to the assigned clustering result, $\boldsymbol{Z}^{mix}$ is separated to $\boldsymbol{Z}^{heart}$ and $\boldsymbol{Z}^{lung}$by deactivating the submatrices which do not belong to the target, respectively.

After obtaining the coded matrix of each source, we decode it as Eqs. (9) and (10).

$$\boldsymbol{Y}^{heart} = D(\boldsymbol{Z}^{heart}) \qquad (9)$$

$$\boldsymbol{Y}^{lung} = D(\boldsymbol{Z}^{lung}). \qquad (10)$$

In the proposed approach, to obtain superior performance of the separated sounds, we first compute the ratio mask of these two sounds, which are defined as Eqs. (11) and (12).

$$\boldsymbol{M}^{heart} = \left(\frac{D(\boldsymbol{Z}^{heart})}{D(\boldsymbol{Z}^{heart}) + D(\boldsymbol{Z}^{lung})} \right) \qquad (11)$$

$$\boldsymbol{M}^{lung} = \left(\frac{D(\boldsymbol{Z}^{lung})}{D(\boldsymbol{Z}^{heart}) + D(\boldsymbol{Z}^{lung})} \right) . \qquad (12)$$

With the estimated $\boldsymbol{M}^{heart}$ and $\boldsymbol{M}^{lung}$, we obtain the heart LPS $\hat{\boldsymbol{Y}}^{heart}$ and lung LPS $\hat{\boldsymbol{Y}}^{lung}$ by Eqs. (13) and (14).

$$\hat{\boldsymbol{Y}}^{heart} = \boldsymbol{M}^{heart} \odot D(\boldsymbol{Z}^{mix}) \qquad (13)$$

$$\hat{\boldsymbol{Y}}^{lung} = \boldsymbol{M}^{lung} \odot D(\boldsymbol{Z}^{mix}), \qquad (14)$$

where $\odot$ denotes the element-wise multiplication. Then $\hat{\boldsymbol{Y}}^{heart}$ and $\hat{\boldsymbol{Y}}^{lung}$ along with the original phase are used to obtain the separated heart and lung waveforms.

IV. Experiments

A. *Experimental setups*

In addition to the proposed PC-DAE(F) and PC-DAE(C), we tested some well-known approaches for comparison, including direct-clustering NMF (DC-NMF), PC-NMF, and deep clustering based on DAE (DC-DAE). The PC-NMF and PC-DAE shared a similar functionality where the PC-DAE performs clustering on the latent representations for heart and lung sound separation. For a fair comparison, the DC-NMF, PC-NMF, and DC-DAE implemented in this study are carried out in an unsupervised manner. For all the methods, the mixed spectrograms were used as the input, and the separated heart and lung sounds were generated at the output.

The DAE(F) model consisted of seven hidden layers, and the neurons in these layers were 1024, 512, 256, 128, 256, 512, and 1024. The encoder of the DAE(C) model consisted of three convolutional layers. The first layer had 32 filters with a kernel size of 1×4, the second layer had 16 filters with a kernel size of 1×3, and the third layer had 8 filters with a kernel size of 1×3 of the encoder. The decoder comprised of four layers. The first layer had 8 deconvolutional filters with a kernel size of 1×3, the second layer had 16 deconvolutional filters with the kernel size of 1×3, the third layer had 32 deconvolutional filters with a kernel size of 1×4, and the fourth layer had 1 deconvolutional filter with kernel size of 1 × 1. Both convolution and deconvolution units adopt a stride of 1. The rectified linear unit were used in encoder and decoder, and the optimizer was Adam. The unsupervised NMF-based methods were used as the baseline, where the basis number of NMF was set to 20, and the L2 norm was used as the cost function. The NMF approach first decomposes the input spectrogram ***V*** into the basis matrix ***W*** and the weight matrix ***H***, where ***W*** serves as the sound basis (including both heart and lung sounds), and ***H*** are the weighting coefficients:

$$V_{ij} \approx (WH)_{ij} = \sum_{a=1}^{r} W_{ia}H_{aj}, \qquad (15)$$

where V_{ij} is the *ij*-th component of ***V*** (a matrix that contains multiple sound sources) and W_{ia} and H_{aj} are the *ia*-th component of ***W*** and the *ai*-th component of ***H***, respectively.

For unsupervised source separation, the weighting coefficient matrix ***H*** is clustered into several distinct groups. When performing separation, the target source of interest can be reconstructed by using the group of vectors in ***H*** that corresponds to the target source. Because the clustering is directly applied to the weighting matrix, we refer to this approach as DC-NMF as the first baseline system. Rather than directly clustering, the PC-NMF [49] clusters the vectors in ***H*** based on the periodicity of individual sound sources; the PC-NMF was also implemented as the second baseline.

Recently, a deep clustering technique [56] that combines a deep learning algorithm and a clustering process has been proposed and confirmed effective for speech [45] and music [46] separation. The fundamental theory of deep clustering is similar to DC-NMF as the clustering is applied in the latent representations instead of the weighting matrix. Because the deep-learning models first transform the input spectrograms into more representative latent features, the clustering of latent features can provide superior separation results. In this study, we implement a deep clustering approach as another comparative method. We used the model architecture of DAE(C) as the deep-learning-based model when implementing the deep clustering approach; hence, the approach is terms DC-DAE(C).

For all the separation methods conducted in this study, we can obtain separated heart and lung sounds. We used the pure heart and lung sounds as a reference to compute the separation performance and adopted three standardized evaluation metrics, namely signal distortion ratio (SDR), signal to interferences ratio (SIR), and signal to artifacts ratio (SAR) [57] to evaluate the separation performances. In a source separation task, there are three types of noise: (1) noise due to missed separation (e_{interf}); noise due to the reconstruction process (e_{artif}), and the perturbation noise (e_{noise}). The computations of SDR, SIR, and SAR are presented in Eqs. (16)-(19), where $\hat{s}(t)$ is the estimated result and $s_{target}(t)$ is the target.

$$\hat{s}(t) = s_{target}(t) + e_{interf} + e_{noise} + e_{artif} \qquad (16)$$

$$\text{SDR} \coloneqq 10\, log_{10} \frac{\left\|s_{target}(t)\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2} \qquad (17)$$

$$\text{SIR} \coloneqq 10\, log_{10} \frac{\left\|s_{target}(t)\right\|^2}{\left\|e_{interf}\right\|^2} \qquad (18)$$

$$\text{SAR} \coloneqq 10\, log_{10} \frac{\left\|s_{target}(t) + e_{interf} + e_{noise}\right\|^2}{\left\|e_{artif}\right\|^2}. \qquad (19)$$

For all of these three metrics, higher scores indicate better source separation results.

We conducted experiments using two datasets. In the first dataset, the heart and lung sounds were collected by SAM, which is a standard equipment in teaching and learning heart and lung sounds[48]. Fig. 5 shows the model of SAM. The SAM attempts to simulate the real human body and has many speakers inside its body corresponding to organ's positions. The SAM can generate clean heart sound or lung sound in different locations. We used the iMEDIPLUS electronic stethoscope [58] to record heart and lung sounds in an anechoic chamber. The heart sounds used in this experiment included normal heart sounds with two beats (S1 and S2). The lung sounds in this experiment included normal, wheezing, rhonchi, and stridor sounds. Both heart and lung sounds were sampled at 8k Hz. The two sounds were mixed at different signal to noise ratio (SNR) levels (-6 dB, -2 dB, 0 dB, 2 dB, and 6 dB) using pure heart sound as the target signal and pure lung heart sound as the noise signal. All the sounds were converted into spectral-domain by applying the short-time Fourier transform (STFT) with a 2048 frame length and 128 frame shifts. Because high frequency parts may not provide critical information for further analyses, we only use 0-300 bins (corresponding to 0-1170 Hz) in this study.

Fig. 5. Student Auscultation Manikin (SAM).

B. Latent space analysis of a selected case

In this section, we used a sample mixed sound to detail every step in the PC-DAE system. Fig. 6 shows the overall procedure of the PC-DAE, where Fig. 6 (a) and (b) show the spectrograms of pure heart and lung sounds, respectively. Fig. 6 (c) shows the latent representation extraction process. For demonstration purpose, we selected two specific neurons, one corresponding to heart sounds and the other corresponding to lung sounds, and plotted their trajectories along the time axis in Fig. 6 (d) and (e), respectively. By evaluating Fig. 6 (d) and (e), we first perceive that the periodicity properties of Fig. 6 (d) and (e) aligned well with Fig. 6 (a) and (b), respectively. Meanwhile, we observe different trajectories of these two neurons, and the periodicity of heart sound is different from lung sound. Next, we applied the DFT on the trajectories of Fig. 6 (d) and (e) and obtained Fig. 6 (f) and (g), respectively, to capture the periodicity more explicitly. Notably, the x-axis for Fig. 6 (a), (b), (d), and (e) is time (s), while the x-axis of Fig. 6 (f) and (g) is frequency (Hz). In the temporal signal analysis, the signals in Fig. 6 (f) and (g) are termed MFA [59] of Fig. 6 (d) and (e) . As can be seen by converting the trajectory into the modulation domain, the periodicity can be more easily observed.

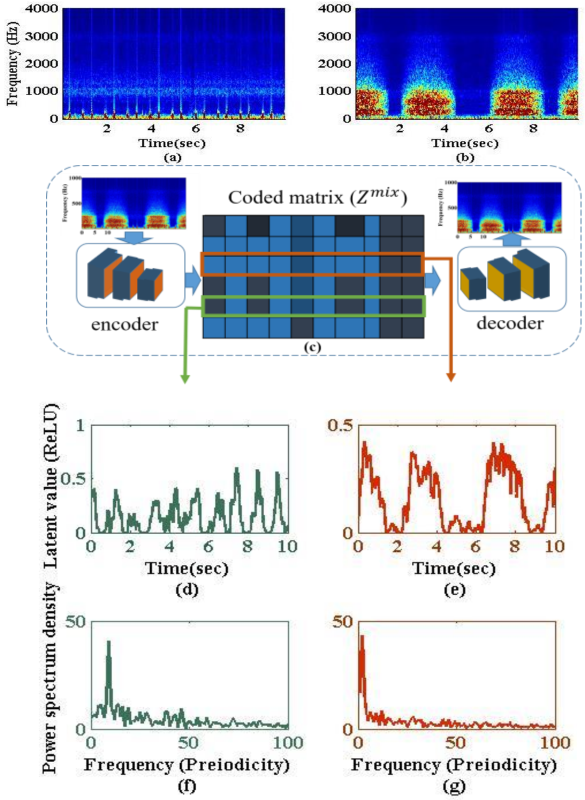
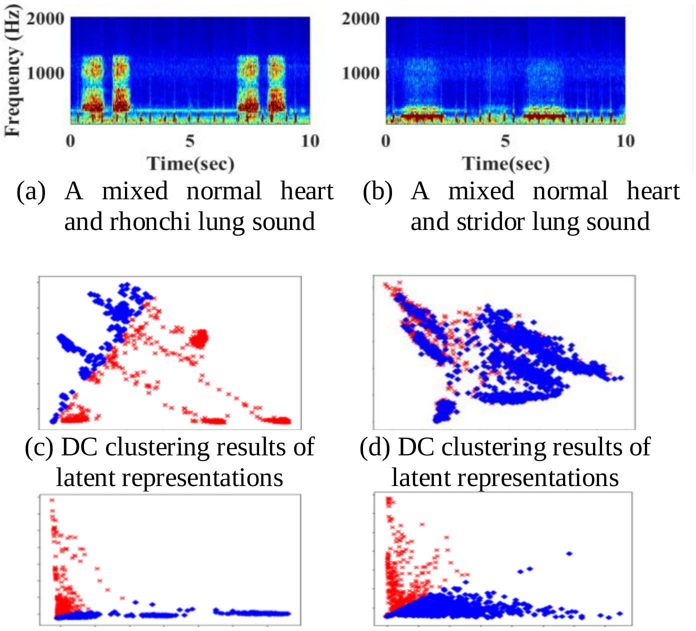


Fig. 6. Analyses of latent representations of sample sounds. (a) and (b), respectively, are the spectrograms of the pure heart and lung sounds, the x-axis is time (s) and y-axis is frequency (Hz); (c) presents the latent representation extraction based on the DAE model; (d) and (e) are trajectories of two latent neurons, where the x-axis is the time, and the y-axis is activation value (0-1); (f) and (g) is the DFT results, where the x-axis is the frequency and y-axis denotes the power density.

By comparing Fig. 6 (f) and (g), we observe a peak in the low-frequency part in Fig. 6 (f), and a peak in Fig. 6 (g) is located at a high-frequency part in Fig. 6 (g). The results suggest that these two neurons should be clustered into two different groups. We apply the same procedures (trajectory extraction and DFT) on all the neurons in the DAE. The neurons that process shorter and longer periodicity are clustered into two distinct groups. Finally, given a mixed sound, we first extract the latent representation; to extract heart sounds, we then keep the neurons that correspond to heart sounds and deactivated the neuron that corresponds to lung sounds and vice versa.

To further verify the effectiveness of the PC clustering approach, we compare DC and PC clustering approaches by qualitatively analyzing the clustering results. To facilitate a clear visual comparison, we adopted the principle component analysis (PCA) [60] to reduce the dimensions on the latent representations to only 2-D and then draw the scattering plots in Fig. 7. The figure shows the spectrograms of two mixed heart-lungs sounds and the clustering results of latent representations. Fig. 7(a) shows the spectrogram of a mixed normal heart sound and abnormal lung (rhonchi) sound; Fig. 7 (b) shows the spectrogram of a mixed normal heart sound and abnormal lung (stridor) sound. Fig. 7 (c) and (d) are the DC clustering results of latent representations (dimensionality-reduced by PCA) corresponding to Fig. 7 (a) and (b), respectively. Fig. 7 (e) and (f) are the PC clustering results of the latent representations corresponding to Fig. 7 (a) and (b), respectively.



(a) A mixed normal heart and rhonchi lung sound

(b) A mixed normal heart and stridor lung sound

(c) DC clustering results of latent representations

(d) DC clustering results of latent representations

(e) PC clustering results of latent representations

(f) PC clustering results of latent representations

Fig. 7. Spectrograms of two mixed heart-lung sounds and the clustering results of latent representations. (a) and (b) are the spectrograms of two mixed heart and lung sounds; (c) and (d)

Table 1
Evaluation results of separated heart sounds generated by the proposed PC-DAE(F) and PC-DAE(C) comparing to three conventional approaches in terms of SDR, SIR, and SAR. Avg denotes the average scores over five SNRs.

	DC-NMF			PC-NMF			DC-DAE(C)			PC-DAE(F)			PC-DAE(C)		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-6dB	-2.74	0.22	5.19	-2.08	1.09	6.09	0.86	2.02	11.24	-1.01	0.32	9.58	2.37	3.57	12.00
-2dB	-0.45	3.76	4.54	-0.92	2.66	6.60	3.01	4.58	10.91	2.36	3.94	10.51	6.46	7.81	13.78
0dB	0.29	5.04	4.61	-1.12	2.17	5.87	5.68	7.60	13.31	3.52	5.96	10.43	7.57	9.17	14.32
2dB	0.81	5.71	4.69	1.75	6.14	6.97	6.35	7.97	14.23	5.59	8.14	10.89	9.38	10.97	15.40
6dB	2.49	9.09	4.70	4.82	11.19	7.88	8.49	11.07	14.04	7.54	11.70	10.46	12.54	14.97	16.79
Avg	0.08	4.76	4.75	0.49	4.65	6.68	4.88	6.65	12.75	3.60	6.01	10.38	8.72	10.44	14.95

Table 2
Evaluation results of separated lung sounds generated by the proposed PC-DAE(F) and PC-DAE(C) comparing to three conventional approaches in terms of SDR, SIR, and SAR. Avg denotes the average scores over five SNRs.

	DC-NMF			PC-NMF			DC-DAE(C)			PC-DAE(F)			PC-DAE(C)		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-6dB	-2.71	-0.01	5.28	-2.48	0.54	5.94	-0.97	0.23	10.68	-1.15	0.11	9.39	3.40	4.64	12.69
-2dB	-0.02	3.62	5.78	-0.52	3.32	5.80	3.11	5.10	11.37	2.70	4.26	10.73	6.94	9.45	13.14
0dB	1.04	4.80	6.00	0.93	6.25	5.45	4.82	6.70	13.12	3.40	5.10	11.25	8.17	10.82	14.11
2dB	2.44	6.89	6.07	1.84	7.58	5.40	5.62	8.50	11.66	5.56	7.86	11.65	9.16	12.05	13.96
6dB	3.34	9.21	6.08	3.34	9.66	6.26	8.50	12.02	12.45	8.30	11.61	12.10	10.90	14.88	14.16
Avg	0.82	4.90	5.84	0.62	5.47	5.77	4.22	6.51	11.86	3.76	5.79	11.02	7.71	10.37	13.61

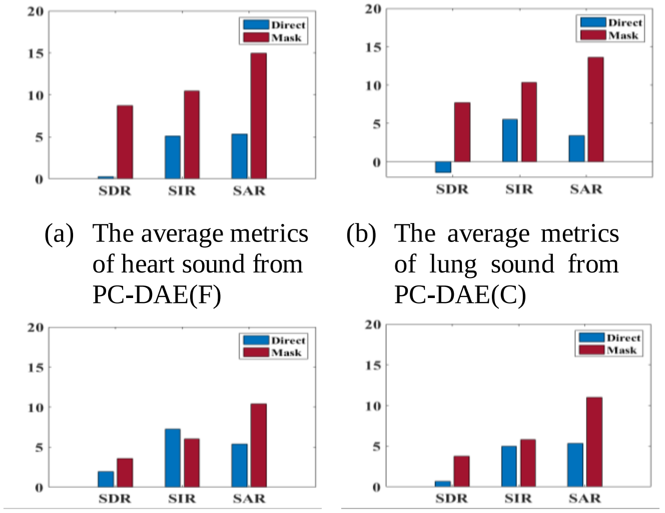
are the DC clustering results of the latent representation; (e) and (f) are the PC clustering results of the latent representation.

By observing Fig. 7 (a), (c), and (e), we can note that heart and lung sounds showed clearly different time-frequency properties (as shown in Fig. 7(a)). In this case, both DC (as in Fig. 7(c)) and PC (as shown in Fig. 7(e)) clustering approaches can effectively group the latent features corresponding to lung and heart sounds into two distinct groups. Consequently, satisfactory separation results can be achieved for both DC and PC approaches. Next, by observing the results of Fig. 7 (b), (d), and (f), since the stridor sound are highly overlapped with heart sound (as show in Fig. 7(b)), the DC clustering approach (as show in Fig. 7(d)) cannot effectively group the latent representations into two distinct groups. On the other hand, the PC clustering approach (as show in Fig. 7(f)) can successfully cluster the latent representations into two distinct groups and consequently yield better separation results.

Please note that any particular time-frequency representation method can be used to perform MFA. The present study adopts the DFT as a representative method. Other time-frequency representation methods, such as CWT [29-31][61] and Hilbert–Huang transform [62-64], can be used. When using these methods, suitable basis functions or prior knowledge need to be carefully considered. In this study, we intend to focus our attention on DFT and will further explore other time-frequency representation methods in the future.

C. Quantitative evaluation based on source separation evaluation metrics

Next, we intend to compare the separation performance using Eqs. (9) and (10) and Eqs. (13) and (14). The results are listed in Fig. 8. Since Eqs. (9) and (10) directly estimate the hear sound and lung sounds, the results using Eqs. (9) and (10) are termed “Direct”. On the other hand, because Eqs. (13) and (14) estimate the heart and lung sounds by a ratio mask function, results are termed “Mask”. We tested the performance using both PC-DAE(F) and PC-DAE(C). From the results in Fig. 8, we observe the results of “Mask” consistently outperform that of “Direct” except for heart sound’s SIR of DAE(F), and confirm the effectiveness of using a ratio mask function to perform separation instead of direct estimation. In the following discussion, we only report the PC-DAE separation results using the ratio mask functions of Eqs. (13) and (14).



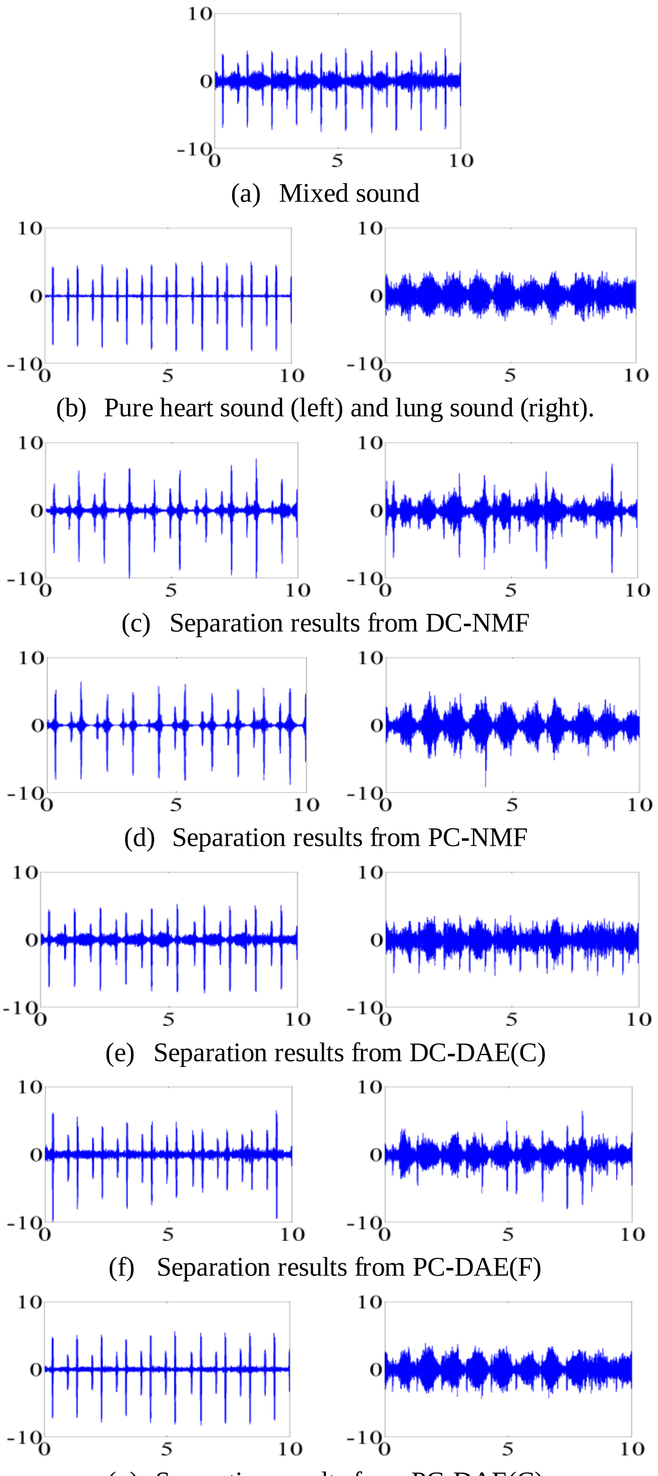
(a) The average metrics of heart sound from PC-DAE(F)

(b) The average metrics of lung sound from PC-DAE(C)

(c) The average metrics of heart sound from PC-DAE(F)

(d) The average metrics of lung sound from PC-DAE(F)

Fig. 8. Average separation results over different SNR conditions. (a) and (c) show the heart sound separation results using PC-DAE(C) and PC-DAE(F), respectively; (b) and (d) show the lung sound separation results using PC-DAE(C) and PC-DAE(F), respectively.



(a) Mixed sound

(b) Pure heart sound (left) and lung sound (right).

(c) Separation results from DC-NMF

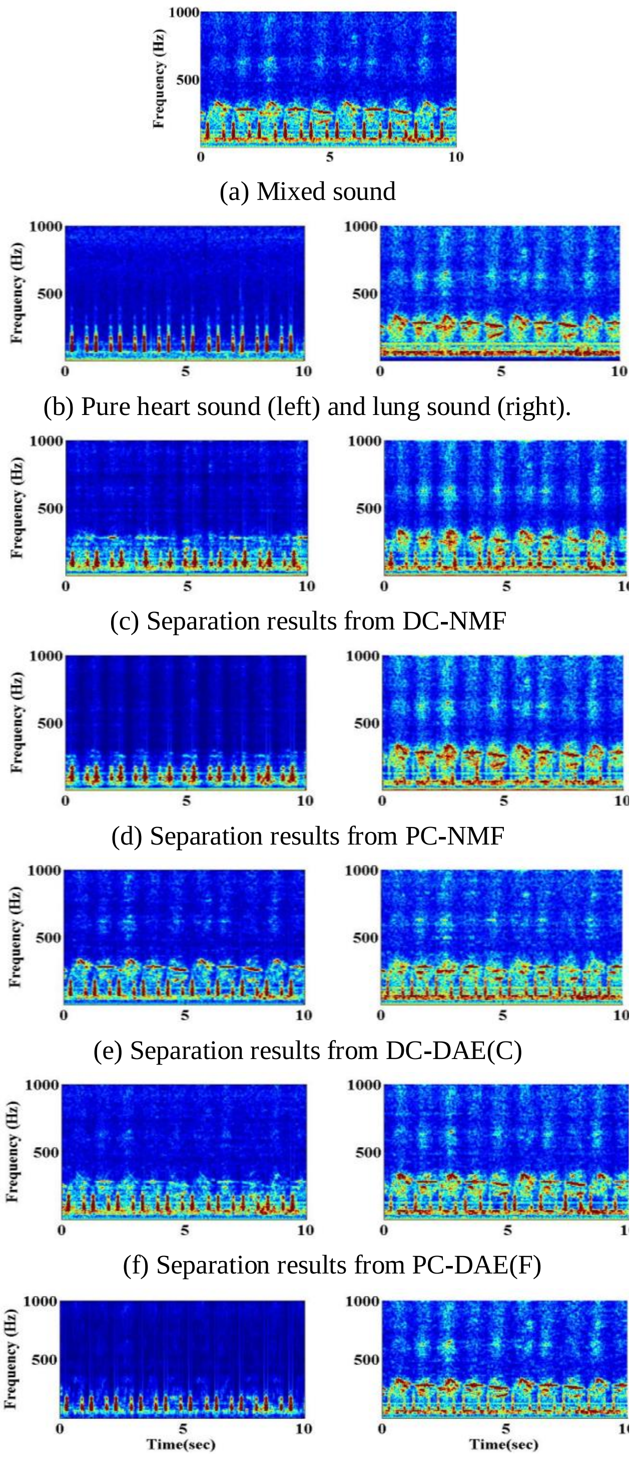
(d) Separation results from PC-NMF

(e) Separation results from DC-DAE(C)

(f) Separation results from PC-DAE(F)

(g) Separation results from PC-DAE(C)

Fig. 9. The waveform of a mixed sample. The y-axis is the amplitude of the signals, and the x-axis is time index (sec). From (b) to (g), the left and right panels are heart sound and lung sound, respectively.



(a) Mixed sound

(b) Pure heart sound (left) and lung sound (right).

(c) Separation results from DC-NMF

(d) Separation results from PC-NMF

(e) Separation results from DC-DAE(C)

(f) Separation results from PC-DAE(F)

(g) Separation results from PC-DAE(C)

Fig. 10. Spectrograms of a mixed sample. The y-axis is the frequency of the signals, and the x-axis is time index (s). From (b) to (g), the left and right panels are heart sound and lung sound, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3016831, IEEE Journal of Biomedical and Health Informatics

Tables 1 and 2 show the evaluation results of heart and lung sounds, respectively, tested on the proposed PC-DAE(F) and PC-DAE(C) with comparative methods. The separation performance is consistent for heart and lung sounds. From the two tables, we observe all the SDR, SIR, and SAR scores mostly increase along with increasing SNR levels. Meanwhile, we note that PC-NMF outperforms DC-NMF, and PC-DAE(C) outperforms DC-DAE(C), confirming the periodicity property to provide superior separation performance than direct clustering. Meanwhile, we observed that the deep learning-based approaches, namely DC-DAE(C) and PC-DAE(C), outperform NMF-based counterparts, namely DC-NMF and PC-NMF, verifying the effectiveness of deep learning models to extract representative features over shallow models. Finally, we observe that PC-DAE(C) outperforms PC-DAE(F), suggesting that the convolutional architecture can yield superior performance than fully connected architecture for this sound separation task.

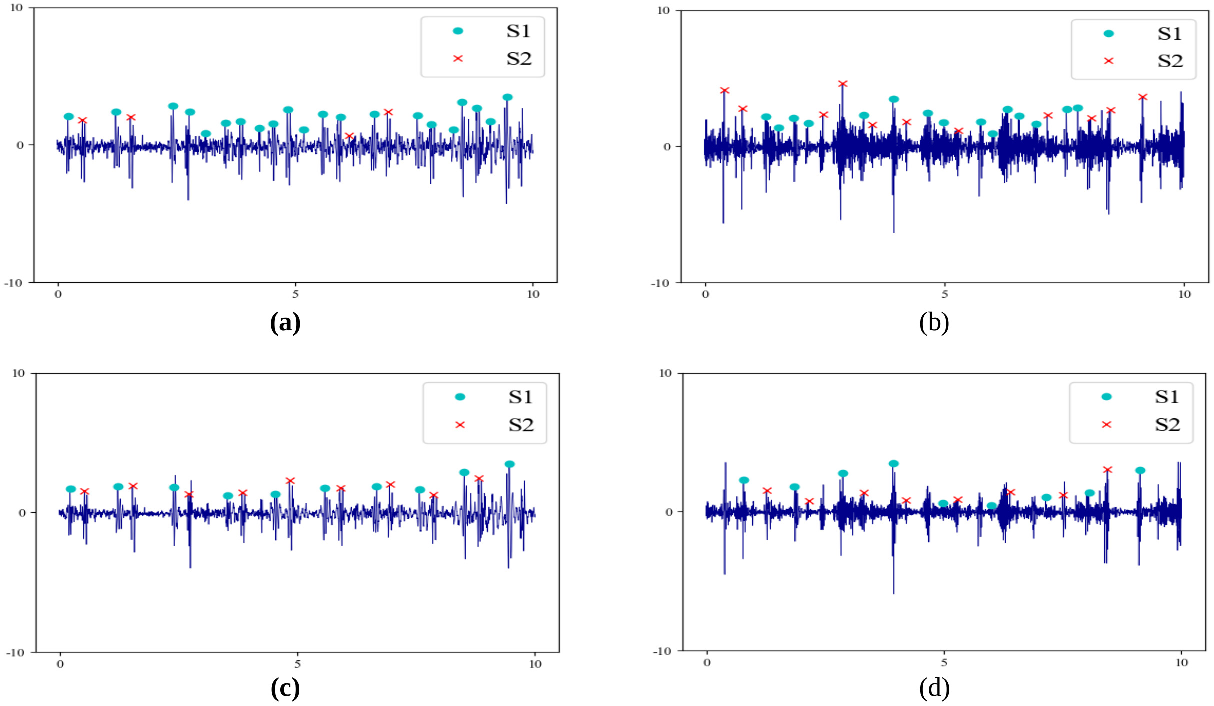
D. Qualitative comparison based on separated waveforms and spectrograms

In addition to quantitative comparison, we also demonstrate waveforms and spectrums of a sample sound to visually compare the separation results. We selected a sample sound, which is the mixed sound with the SNR ratio of heart sound (treated as the signal) and wheezing lung sound (treated as the noise) to be 6 dB. Fig. 9 demonstrates the waveforms of the sample sound, where Fig. 9(a) shows the mixed sounds. Fig. 9(b) shows the pure heart sound (left panel) and lung sound (right panel) that have not been mixed. Fig. 9(c), (d), (e), (f), and (g) show the separated results of DC-NMF, PC-NMF, DC-DAE(C), PC-DAE(F), and PC-DAE(C), respectively. From Fig. 9, we observe that PC-DAE(C) can more effectively separate the heart and lung sounds as compared to other methods; the trends are consistent with those shown in Tables 1 and 2.

Next in Fig. 10, we show the spectrograms of the same sample sound shown in Fig. 9. Fig. 10(a) presents the mixed sounds, Fig. 10(b) shows the pure heart and lung sounds, and Fig. 10(c) to (g) are separated results corresponding to Fig. 10(c) to (g). From Fig. 10(a), we can observe that the two sounds are highly overlapped in the lower frequency region. It is also noticed that PC-NMF possesses a higher performance for interference suppression during the high frequency of lung sounds, and PC-DAE(F) possesses a higher performance in overlapped frequency bandwidth and receives improved heart sound quality. PC-DAE(F) and PC-DAE(C) performed the best with minimal

Table 3
Recognition accuracies of mixed heart-lung sounds and separated heart sounds with different age and gender groups.

	Mixed heart-lung sound Separated						Separated heart sound					
Age	0-20		21-65		66-80		0-20		21-65		66-80	
Gender	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Accuracy	71%	67%	67%	76%	71%	57%	81%	90%	80%	85%	86%	90%
Avg	69%		72%		64%		86%		83%		88%	



(a) (b)

(c) (d)

Fig 11. The waveforms of two sound samples and the corresponding S1-S2 recognition results. (a) a mixed heart-lung sound with normal heart sound and normal lung sound. (b) a mixed heart-lung sound with abnormal heart sound and abnormal lung sound. (c) and (d) are the separated results corresponding to (a) and (b), resepctively. The recognized S1 and S2 results are colored by green and red symbols, respectively.

artificial noises. Generally speaking, the two PC-DAE approaches outperformed the other approaches yielding clear separation spectrograms.

E. Real application in first heart sound (S1) and second heart sound (S2) recognition

We used another dataset to further evaluate the proposed algorithm in a more real-world scenario. Real mixed heart-lung sounds were collected from National Taiwan Hospital, and the proposed PC-DAE was used to separate the heart and lung sounds. Because it is not possible to access pure heart and lung sounds corresponding to the mixed heart-lung sounds, the SDR, SIR, and SAR scores cannot be used as the evaluation metrics in this task. Instead, we adopted the first heart sound (S1) and second heart sound(S2) recognition metric accuracies to determine the separation performance. We adopted a well-known S1 and S2 recognition algorithm from [10, 65], which considers frequency properties and the assumption of S1-S2 and S2-S1 intervals. We believe that this alternative metric is convincing and valuable since the S1-S2 recognition accuracy has already been used as a crucial index for doctors to diagnose the occurrence of diseases[66, 67].

This dataset includes 3 different age groups, namely 0-20 (childhood and adolescence), 21-65 (adulthood), and over 66 (senior citizen)). Each group has 6 cases, including 3 males and 3 females, and each case has 7 mixed heart-lung sounds (10 sec). Based on this design, we can determine whether the proposed approach can be robust against variations of age and gender groups (accordingly covering people with different physiological factors, such as blood pressure, heart rate, etc.). Table. 3 shows the recognition accuracies of before and after performing heart-lung sound separation.

To visually investigate the S1-S2 recognition performance, we present the waveforms along with the recognition results in Fig. 11. Fig 11 (a) and (b) are two sound samples, where Fig. 11 (a) is the mixed heart-lung sound with normal heart and lung sounds, and Fig. 11 (b) is the mixed heart-lung sound with abnormal heart sound (weak periodicity) and abnormal lung sound (rhonchi). Fig 11 (c) and (d) show the S1-S2 recognition the after performing heart-lung sound separation corresponding to Fig 11 (a) and (b), respectively.

From Fig. 11 (a) and (b), we can note that the S1-S2 recognition results are poor for the mixed sounds, and the recognition performance are notably improved with the separated heart sounds (as can be seen from Fig 11 (c) and (d)), confirming the effect of the proposed PC-DAE's outstanding capability of separating the heart sounds from mixed sounds.

V. Conclusion

The proposed PC-DAE is derived based on the periodicity properties of the signal to perform blind source separation in a single-channel recording scenario. Different from the conventional supervised source separation approach, PC-DAE does not require supervised training data. To the best of our knowledge, the proposed PC-DAE is the first work that combines the advantages of deep-learning-based feature representations and the periodicity property to carry out heart-lung sound separations. The results of this study indicate that the proposed method is effective to use a periodic analysis algorithm to improve the separation of sounds with overlapped frequency bandwidth. The results also show that PC-DAE provided satisfactory separation results and achieve superior quality as compared to several related works. Moreover, we verified that by using the proposed PC-DAE as a preprocessing step, the heart sound recognition accuracies can be considerably improved. In our current work, we need to define how many sources are in the signal. However, in most cases, determining the exact number of the sources is difficult. Hence, identifying on effective way to determine the number of the sources is an important future work. In the present study, we consider the condition where only sounds recorded by an electronic stethoscope is available. We believe that this experiment setup is close to most real-world clinical scenarios. In the future, we will extend the proposed PC-DAE to the conditions where additional physiological data is available, such as ECG, photoplethysmogram, and blood pressure signals.

Reference

[1] A. B. Bohadana, R. Peslin, H. Uffholtz, and G. Pauli, "Potential for lung sound monitoring during bronchial provocation testing," *Thorax,* vol. 50, no. 9, pp. 955-961, 1995.

[2] J. Hardin and J. Patterson Jr, "Monitoring the state of the human airways by analysis of respiratory sound," *Acta Astronautica,* vol. 6, no. 9, pp. 1137-1151, 1979.

[3] C. Ahlstrom *et al.*, "Feature extraction for systolic heart murmur classification," *Annals of biomedical engineering,* vol. 34, no. 11, pp. 1666-1677, 2006. [Online]. Available: https://link.springer.com/article/10.1007/s10439-006-9187-4.

[4] A. Jones, "A brief overview of the analysis of lung sounds," *Physiotherapy,* vol. 81, no. 1, pp. 37-42, 1995.

[5] J. Cummiskey, T. C. Williams, P. E. Krumpe, and C. Guilleminault, "The detection and quantification of sleep apnea by tracheal sound recordings," *American Review of Respiratory Disease,* vol. 126, no. 2, pp. 221-224, 1982.

[6] G. Limongelli and B. Sarubbi, "Atrial arrhythmias in adults with congenital heart disease. Listening to your heart sound can save your life," *International journal of cardiology,* vol. 248, pp. 159-160, 2017.

[7] M. Silvestri, S. Franchi, A. Pistorio, L. Petecchia, and F. Rusconi, "Smoke exposure, wheezing, and asthma development: A systematic review and meta-analysis in unselected birth cohorts," *Pediatric pulmonology,* vol. 50, no. 4, pp. 353-362, 2015.

[8] S. Choi and Z. Jiang, "Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique," *Computers in biology and medicine,* vol. 40, no. 1, pp. 8-20, 2010.

[9] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Computers in Cardiology,* 1997.

[10] D. Kumar *et al.*, "Detection of S1 and S2 heart sounds by high frequency signatures," in *International Conference of the IEEE Engineering in Medicine and Biology Society,* 2006.

[11] J. E. Hebden and J. Torry, "Neural network and conventional classifiers to distinguish between first and second heart sounds," 1996.

[12] A. C. Stasis, E. Loukis, S. Pavlopoulos, and D. Koutsouris, "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system," in *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine,* 2003.

[13] M. Bahoura, "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes," *Computers in biology and medicine,* vol. 39, no. 9, pp. 824-843, 2009.

[14] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial intelligence in medicine,* vol. 88, pp. 58-69, 2018.

[15] S. Abbasi, R. Derakhshanfar, A. Abbasi, and Y. Sarbaz, "Classification of normal and abnormal lung sounds using neural network and support vector machines," in *Iranian Conference on Electrical Engineering (ICEE)*, 2013.

[16] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in adaptive data analysis,* vol. 1, no. 01, pp. 1-41, 2009.

[17] J. K. Peat, R. H. Van Den Berg, W. F. Green, C. M. Mellis, S. R. Leeder, and A. Wolcock, "Changing prevalence of asthma in Australian children," *Bmj,* vol. 308, no. 6944, pp. 1591-1596, 1994.

[18] N. Gavriely, M. Nissan, D. Cugell, and A. Rubin, "Respiratory health screening using pulmonary function tests and lung sound analysis," *European Respiratory Journal,* vol. 7, no. 1, pp. 35-42, 1994.

[19] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare," *IEEE journal of translational engineering in health and medicine,* vol. 6, pp. 1-12, 2018.

[20] D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali, "Computerized lung sound screening for pediatric auscultation in noisy field environments," *IEEE Transactions on Biomedical Engineering,* vol. 65, no. 7, pp. 1564-1574, 2017.

[21] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine, and W. Checkley, "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis," *Respiratory medicine,* vol. 105, no. 9, pp. 1396-1403, 2011.

[22] V. K. Iyer, P. Ramamoorthy, H. Fan, and Y. Ploysongsang, "Reduction of heart sounds from lung sounds by adaptive filterng," *IEEE transactions on biomedical engineering,* no. 12, pp. 1141-1148, 1986.

[23] Z. Wang, J. N. da Cruz, and F. Wan, "Adaptive Fourier decomposition approach for lung-heart sound separation," in *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2015.

[24] J. Gnitecki, Z. Moussavi, and H. Pasterkamp, "Recursive least squares adaptive noise cancellation filtering for heart sound reduction in lung sounds recordings," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, 2003.

[25] L. J. Hadjileontiadis and S. M. Panas, "Adaptive reduction of heart sounds from lung sounds using fourth-order statistics," *IEEE Transactions on Biomedical Engineering,* vol. 44, no. 7, pp. 642-648, 1997.

[26] L. Yip and Y. Zhang, "Reduction of heart sounds from lung sound recordings by automated gain control and adaptive filtering techniques," in *Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2001.

[27] C. Lin, W. A. Tanumihardja, and H. Shih, "Lung-heart sound separation using noise assisted multivariate empirical mode decomposition," in *International Symposium on Intelligent Signal Processing and Communication Systems*, 2013.

[28] A. Mondal, P. Bhattacharya, and G. Saha, "Reduction of heart sound interference from lung sound signals using empirical mode decomposition technique," *Journal of medical engineering & technology,* vol. 35, no. 6-7, pp. 344-353, 2011.

[29] I. Hossain and Z. Moussavi, "An overview of heart-noise reduction of lung sound using wavelet transform based filter," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, 2003.

[30] L. J. Hadjileontiadis and S. M. Panas, "A wavelet-based reduction of heart sound noise from lung sounds," *International Journal of Medical Informatics,* vol. 52, no. 1-3, pp. 183-190, 1998.

[31] M. Pourazad, Z. Moussavi, and G. Thomas, "Heart sound cancellation from lung sound recordings using time-frequency filtering," *Medical and biological engineering and computing,* vol. 44, no. 3, pp. 216-225, 2006.

[32] F. Ayari, M. Ksouri, and A. T. Alouani, "Lung sound extraction from mixed lung and heart sounds FASTICA algorithm," in *IEEE Mediterranean Electrotechnical Conference*, 2012.

[33] J.-C. Chien, M.-C. Huang, Y.-D. Lin, and F.-c. Chong, "A study of heart sound and lung sound separation by independent component analysis technique," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006.

[34] B. Makkiabadi, D. Jarchi, and S. Sanei, "A new time domain convolutive BSS of heart and lung sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[35] C. Lin and E. Hasting, "Blind source separation of heart and lung sounds based on nonnegative matrix factorization," in *International Symposium on Intelligent Signal Processing and Communication Systems*, 2013.

[36] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing,* vol. 15, no. 3, pp. 1066-1074, 2007.

[37] H. Wang, "Multi-sensor adaptive heart and lung sound extraction," in *SENSORS, IEEE*, 2003,.

[38] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[39] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[40] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International conference on latent variable analysis and signal separation*, 2017.

[41] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 9, pp. 1652-1664, 2016.

[42] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 12, pp. 2136-2147, 2015.

[43] T. H. Lin and Y. Tsao, "Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval," *Remote Sensing in Ecology and Conservation*, 2019.

[44] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, 2013.

[45] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[46] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[47] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication,* vol. 28, no. 1, pp. 43-55, 1999.

[48] Cardionics. "SAM® 3G — Student Auscultation Manikin." https://www.cardionics.com/product/learning-systems/sam-3g-student-auscultation-manikin-3rd-generation (accessed.

[49] T.-H. Lin, S.-H. Fang, and Y. Tsao, "Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings," *Scientific reports,* vol. 7, no. 1, p. 4547, 2017.

[50] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering,* vol. 64, no. 11, pp. 2584-2594, 2017.

[51] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," 2010.

[52] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," in *Interspeech*, 2016.

[53] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 26, no. 9, pp. 1570-1584, 2018.

[54] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*: Elsevier, 1992, pp. 65-93.

[55] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," Georgia Institute of Technology, 2008.

[56] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoencoders," in *International conference on neural information processing*, 2017.

[57] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing,* vol. 14, no. 4, pp. 1462-1469, 2006.

[58] Imediplus. "electronic stethoscope DS301." http://www.imediplus.com/ (accessed.

[59] T. H. Falk and W.-Y. Chan, "Modulation filtering for heart and lung sound separation from breath sound recordings," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008.

[60] G. H. Dunteman, *Principal components analysis* (no. 69). Sage, 1989.

[61] S.-S. Wang, P. Lin, Y. Tsao, J.-W. Hung, and B. Su, "Suppression by selecting wavelets for feature compression in distributed speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 3, pp. 564-579, 2017.

[62] Z. Peng, W. T. Peter, and F. Chu, "An improved Hilbert–Huang transform and its application in vibration signal analysis," *Journal of sound and vibration,* vol. 286, no. 1-2, pp. 187-205, 2005.

[63] Y.-L. Tseng, P.-Y. Ko, and F.-S. Jaw, "Detection of the third and fourth heart sounds using Hilbert-Huang transform," *Biomedical engineering online,* vol. 11, no. 1, p. 8, 2012.

[64] N. E. Huang, *Hilbert-Huang transform and its applications*. World Scientific, 2014.

[65] D. Kumar, P. Carvalho, M. Antunes, P. Gil, J. Henriques, and L. Eugenio, "A new algorithm for detection of S1 and S2 heart sounds," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[66] D. Sapire, "Understanding and diagnosing pediatric heart disease: Heart sounds and murmurs," *Norwalk, Connecticut, Applcton & Langc,* pp. 27-43, 1992.

[67] Y.-N. Wen, A. P.-W. Lee, F. Fang, C.-N. Jin, and C.-M. Yu, "Beyond auscultation: Acoustic cardiography in clinical practice," *International journal of cardiology,* vol. 172, no. 3, pp. 548-560, 2014.

Kun-Hsi Tsai received the B.S. degree in the school of medicine from China Medical University, Taichung, Taiwan in 2003 and Ph.D. degree in Department of Biological Sci-ence and Technology from National Chiao Tung University, Hsinchu, Taiwan, in 2015. He is currently a president in Imediplus Inc. His research interests include clinical studies in heart diseases/respiratory diseases, auxiliary diagnosis, and bio-sound processing.

Wei-Chien Wang received the M.S degree in Institute of Computer and Communication engineering from National Cheng Kung University, Tainan, Taiwan. He is currently an algorithm engineer in Imediplus Inc. From 2016 to 2017, he was a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include signal processing, bio-signal enhancement, bio-signal source separation, algorithm development, machine learning, and deep learning.

Chui-hsuan Cheng received the B.S. degree in the school of medicine from China Medical University, Taichung, Taiwan in 2003, and M.S. degree in Graduate Institute of Injury Prevention and Control from Taipei Medical University, Taipei, Taiwan, in 2014, respectively. He received Emergency Medicine Specialist training in Emergency Department of China Medical University, Taichung, Taiwan, from 2005 to 2009. He is currently a Medical Director in Imediplus Inc. His research interests include clinical studies about the signal analysis of ECG and heart / lung sounds.

Chan-Yen Tsai received the B.S. degree in Science in Laboratory Medicine and Biotechnology and M.S. degree in Human Genetics from Tzu Chi University, Hualien, Taiwan, in 1999 and 2001, respectively. He received Ph.D. degree in Microbiology and Immunology from National Yang Ming University, Taipei, Taiwan, in 2013. He is currently a Medical Analyst in Imediplus Inc. His research interests include clinical study design, numerical analysis, and research in human heart/lung sound.

Jou-Kou Wang received the M.D. degree and Ph.D. degree from National Taiwan University, Taipei, Taiwan, in 1980 and 1991 respectively. He completed Residency and fellowship training in the Department of Pediatrics, Nation Taiwan University Hospital, Taipei, Taiwan. He is currently a professor of Pediatrics, and chief of Pediatric Cardiology of National Taiwan University Hospital. His research interests include interventional therapies for structural heart disease and echocardiography.

Tzu-Hao Lin received his Ph.D. degree in Ecology from National Taiwan University, Taipei, Taiwan, in 2013. He is currently an Assistant Research Fellow in Biodiversity Research Center, Academia Sinica. His research interests include bioacoustics, machine learning, and ecological interactions.

Shih-Hau Fang (M'07-SM'13) is a Full Professor in the Department of Electrical Engineering, Yuan Ze University (YZU), and MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan. He is currently technical advisor to PTCom Technology Company Ltd., and serves as YZU President's Special Assistant. Prof. Fang's research interests include indoor positioning, mobile computing, machine learning, and signal processing. He is a senior member of IEEE. Prof. Fang has received several awards for his research work, including the Young Scholar Research Award (YZU, 2012), Project for Excellent Junior Research Investigators (MOST, 2013), Outstanding Young Electrical Engineer Award (the Chinese Institute of Electrical Engineering, 2017), Outstanding Research Award (YZU, 2018), Best Synergy Award (Far Eastern Group, 2018), Future Technology Award (MOST, 2019), National Innovation Award (RBMP, 2019) and Y.Z Outstanding Professor (Y.Z. Hsu Science and Technology Memorial Foundation, 2019).

Li-Chin Chen received her Ph.D. degree in Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University. She has receives several awards, including the Xin-Miao Zhuo Information Student Leadership Award of Institute of Information and Computer Machinery (2010), president dducational award of ministry of education (2011), outstanding paper award of graduate student of Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University (2014), and the annual best Ph.D. dissertation awards of graduate institute of biomedical Electronics and Bioinformatics of National Taiwan University (2014). Currently, she is a research specialist in the Research Center for Information Technology Innovation of Academia Sinica, Taiwan. Her research interests focused on informatic in healthcare, including the adoption of informatics application and big data research. Recently, the application focused on using deep learning technology to assist metabolic disease monitoring, cancer treatment optimization, prognosis prediction, and policy making.

Yu Tsao received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently an Associate Research Fellow, Deputy Director, and Director of Artificial Intelligence Computing Center with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include speech and speaker recognition, acoustic and language modeling, audio coding, and bio-signal processing. He is currently an Associate Editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE Signal Processing Letters, and IEICE transactions on Information and Systems. Dr. Tsao received the Academia Sinica Career Development Award in 2017, National Innovation Awards in 2018 and 2019, and Outstanding Elite Award, Chung Hwa Rotary Educational Foundation 2019-2020.