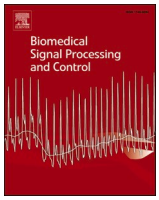
Biomedical Signal Processing and Control 79 (2023) 104175

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Cardiopulmonary auscultation enhancement with a two-stage noise cancellation approach

Chunjian Yang [a], Neng Dai [b,c], Zhi Wang [d,e], Shengsheng Cai [f], Jiajun Wang [a], Nan Hu [a,*]

[a] School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu 215006, China
[b] Department of Cardiology, Zhongshan Hospital, Fudan University, Shanghai Institute of Cardiovascular Diseases, Shanghai 200032, China
[c] National Clinical Research Center for Interventional Medicine, Shanghai 200032, China
[d] State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
[e] Center for Intelligent Acoustics and Signal Processing, Huzhou Institute of Zhejiang University, Huzhou 313000, China
[f] Suzhou Melodicare Medical Technology Co., Ltd., Suzhou 215151, China

ARTICLE INFO

Keywords:
Electronic stethoscope
Cardiopulmonary auscultation
Ambient sound interference
Adaptive noise canceller
Deep neural networks

ABSTRACT

For cardiopulmonary auscultation using electronic stethoscopes, signal quality is a key point. During signal acquisition various background sounds may be inevitably captured, severely degrading the auscultation signal quality. In the existing auscultation denoising methods, conventional adaptive noise canceller (ANC) approaches or shallow-layer artificial neural networks were used, while advanced noise cancellation methods still need to be developed to fulfill real auscultation requirements. In this paper, we propose a novel denoising method for cardiopulmonary auscultation enhancement, which is a two-stage approach with a cascade of ANC and deep neural networks (DNNs). In the first stage, the ANC provides coarsely denoised auscultation signal and estimated interference. In the second stage, a DNN termed the dual-channel interactive noise cancellation network (DINC-Net) is proposed, which exploits both the coarsely denoised auscultation signal and the estimated interference. The DINC-Net consists of two deep encoders extracting features of dual-channel inputs separately, one dual-channel interactive denoising module generating a denoising mask, and one deep decoder giving the denoised output. The performance of the proposed method is evaluated through synthetic data generated using two public heart/lung sound databases, and the great promotion in normalized covariance measure (NCM) and frequency-weighted segmental signal-to-noise ratio (fwSNRseg) has been verified, compared to the existing methods. An online noise cancellation prototype is further developed on an electronic stethoscope, and the signal quality promotion is shown on healthy subjects as well as aortic stenosis patients.

1. Introduction

The electronic stethoscope, as a non-invasive diagnostic instrument, has the advantage of flexibly accessing, recording, and analyzing the physiological acoustic signals from the human body, including cardiac and respiratory sounds [1]. Physicians can assess the cardiopulmonary status of a patient by analyzing the information embedded in auscultation signals via various means. How to aquire high-quality auscultation signal is of great importance in designing an electronic stethoscope, while in real applications various unpredictable interferences may severely deteriorate the auscultation quality [2]. Among these interferences, the ambient noises, widely existing in physician's or pediatrician's offices, occupied the principle position in the causes of auscultation signal corruption.

As noise contamination would limit the clinical application value of auscultation signals, it is essential to develop and integrate an efficient noise cancellation algorithm in the electronic stethoscope. Several monaural denoising methods have been proposed to address this problem [3–8]. Wavelet-domain analysis was involved in [3], and thresholds were set to attenuate wavelet coefficients corresponding to noise for cardiac sound signal enhancement. Empirical mode decomposition (EMD) was utilized in [4] for denoising and analyzing electrocardiogram (ECG) and cardiac sound signals. A short-time Fourier transform (STFT) based semi-automatic filtering technique was introduced in [5] for attenuating noise ingredients from cardiac sound signals. In [6], the study showed that ensemble EMD (EEMD) achieved better performance than wavelet analysis when applied in denoising cardiac sound signals. A combination of wavelet packet transform (WPT) and singular value

* Corresponding author.
E-mail address: hunan@suda.edu.cn (N. Hu).

https://doi.org/10.1016/j.bspc.2022.104175
Received 29 January 2022; Received in revised form 3 August 2022; Accepted 4 September 2022

decomposition (SVD) scheme was proposed to eliminate noise in cardiac sound analysis and validation was performed on a normal/pathological cardiac sound dataset [7]. An integration of EMD, Hurst analysis and spectral subtraction was used to denoise the respiratory sound in [8], and the used data set included 30 chronic obstructive pulmonary disease (COPD) and 30 healthy lung sounds.

The monaural auscultation denoising methods may show advantages in promoting cardiac sound or respiratory sound signal quality, while there are still two issues left to be addressed. First, these studies aimed at denoising either cardiac sound or respiratory sound, by employing the empirical knowledge of the difference between objective signal and noise at some features. However, in realtime implementations the signal to be denoised is usually not single-modal, *e.g.* a mixture of cardiac sound and respiratory sound, and unpredictable and complicated patterns of pathological cardiopulmonary sounds may increase the difficulty of denoising. Second, due to the variability of environments, it's hard to grasp and utilize all information of nonstationary ambient noises via only monaural processing, where auscultation signals are contaminated by ambient noises at time domain as well as at frequency domain. Introducing an extra reference recording of ambient noise picked up by an auxiliary microphone was a promising choice to address the issues mentioned above. The raw recording at the main channel of electronic stethoscope can be deemed as a mixture of pure auscultation signal with noise component in an unknown way, where the mixed noise component is the result of passing the ambient noises recorded by the auxiliary microphone through an unknown system.

As a class of adaptive filter, adaptive noise cancellers (ANCs) were conventional choices for auscultation denoising on two-microphone setup. ANC can adaptively identify the unknown system and then perform noise suppression. The classical least mean square (LMS) algorithm was introduced to enhance lung sound signals in [9]. As normalized least mean square (NLMS) algorithm [10] can provide both fast convergence speed and low computational load, it was applied in [11] to obtain purified lung sounds, where cardiac sound was deemed as noise to be suppressed. In [12], recursive least squares (RLS) ANC was employed to reduce cardiac sound from lung sound recordings. In [13], a spectral subtraction algorithm was designed at frequency domain to improve the auscultation signals' quality.

ANC or spectral subtraction can be deemed as a single-stage denoising approach, in which only a simple linear system is assumed for modeling the unknown system and the manner of signal mixture. Intuitively such a single-stage method, with limited mapping capability, may not handle the complicated denoising problem well in a highly nonstationary noise environment. Since data-driven deep-layered artificial neural networks (ANNs) can automatically learn high-level representations from the input signal, numerous deep-learning-based methods are emerging for the analysis of biomedical signals [14–17]. For instance, transfer learning [18] was used to address the low resource problem of biomedical signals. Based on well trained deep learning models, it requires only a small size of dataset for fine tuning the high layers of the trained models to achieve excellent performance in specific tasks, *e.g.* medical image processing and analysis [19]. More recently, the advantages of multi-stage methods over single-stage methods in many tasks have been revealed [20,21], where data-driven ANNs were involved with ANCs for speech or acoustic signal enhancement. The similar idea was introduced into the task of separating cardiac sound signal from lung sound signal, where a fully connected ANN with one hidden layer was integrated into an adaptive line enhancer (ALE) system [22]. Nevertheless, it is still interesting to study how the multi-stage approach would perform at auscultation signal denoising, and learn the suitable architecture of ANN for this problem.

In this paper, we propose a novel noise cancellation method for cardiopulmonary auscultation enhancement, which is a two-stage approach containing a cascade of conventional ANC and deep neural networks (DNNs). The first stage consists of an ANC based on the NLMS algorithm (ANC-NLMS), which is used to provide coarsely denoised auscultation signal and estimated interference. For the second stage, a DNN named dual-channel interactive noise cancellation network (DINC-Net), is established to eliminate residual noise and distortion. The design of DINC-Net is inspired by a state-of-the-art speech separation solution, the end-to-end fully convolutional time-domain audio separation network (Conv-TasNet) [23], as noise cancellation in electronic stethoscope can be deemed as separating clean cardiopulmonary sounds from noises. Compared to the Conv-TasNet, the proposed DINC-Net builds two deep encoders to extract features of dual-channel inputs, one dual-channel interactive denoising module to generate a denoising mask, and one deep decoder to give an ultimate denoised output. The performance of the proposed method is evaluated via simulation and real application experiments. Synthetic noisy auscultation data are generated from three public datasets: one cardiac sound dataset, one respiratory sound dataset, and one noise dataset, and outperformance of our method is illustrated at different denoising difficulties. An online noise cancellation prototype based on our proposed method is applied in the electronic stethoscope Smartho-D2 [24], and real experiments on healthy subjects and aortic stenosis patients show that the proposed method significantly promote the auscultation signals' quality.

The remainder of this paper is organized as follows. The proposed noise cancellation method and technique details are described in Section 2. Section 3 presents the experiment settings and Section 4 gives the results. Finally, discussions and conclusions are given in Section 5 and Section 6, respectively.

2. Methods

The auscultation denoising method is developed based on a two-microphone setup. The piezoelectric microphone on the primary channel collects the noisy cardiopulmonary auscultation signal $d(t)$, and the auxiliary microphone on the reference channel picks up the ambient noise $v(t)$. After active filtering and analog–digital conversion, the digitalized noisy cardiopulmonary auscultation signal $d(n)$ on the primary channel is given by

$$d(n) = f(s(n),\ h(n) * v(n)) \tag{1}$$

where $s(n)$ is the pure cardiopulmonary sound signal, $h(n)$ is an unknown system function, $*$ denotes convolution operator, and $f(\cdot)$ denotes the unknown function that mixes the pure auscultation signal and noise to form $d(n)$. The auscultation denoising objective is to output a denoised $\widehat{s}(n)$, with inputs $d(n)$ and $v(n)$.

Our proposed auscultation denoising method includes two stages. In the first stage, an ANC based on NLMS is used to give coarsely denoised auscultation signal $s_{\mathrm{ANC}}(n)$ and estimated interference $y(n)$. Due to the limitation in mapping capability of ANC for auscultation denoising, the second stage with a built DINC-Net, whose inputs are $s_{\mathrm{ANC}}(n)$ and $y(n)$, is used to further attenuate the noise, reserve the informative ingredients of useful signal, and finally give the prediction of pure auscultation signal $\widehat{s}(n)$. The DINC-Net is believed to be endowed with the ability of automatically managing the underlying mechanism of noise transfer and signal mixture, by training with datasets considering various real scenarios. The main signal processing flowchart of the proposed method is displayed in Fig. 1.

2.1. The first denoising stage: ANC

The first stage ANC is a natural choice when two inputs are used for denoising [25]. The two inputs are the primary recording $d(n)$ and the reference recording $v(n)$, and the outputs are estimated interference $y(n)$ and coarsely denoised auscultation signal $s_{\mathrm{ANC}}(n)$, given by

$$y(n) = \widehat{h}(n) * v(n) \tag{2}$$

$$s_{\mathrm{ANC}}(n) = d(n) - y(n) \tag{3}$$

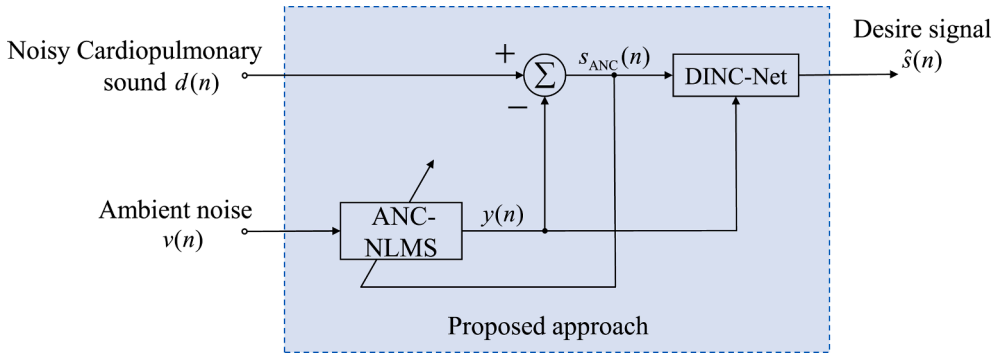


Fig. 1. A block diagram of the proposed two-stage noise cancellation system. ANC-NLMS is used to provide coarsely denoised auscultation signal and estimated interference. The Cascaded DINC-Net is established to further eliminate residual noise and distortion.

where $\widehat{h}(n)$ denotes the transfer function of employed adaptive filter estimating the unknown system.

In this paper, a K-order finite impulse response (FIR) filter is used for $\widehat{h}(n)$ and NLMS is employed to adaptively update the parameters of FIR filter. Hence $y(n)$ is given by

$$y(n) = \widehat{h}(n) * v(n) = \boldsymbol{w}(n)^{\mathrm{T}}\boldsymbol{v}(n) \tag{4}$$

where $\boldsymbol{w}(n)$ is the K-element weight vector estimated for the K-order FIR filter, $(\cdot)^{\mathrm{T}}$ denotes transpose, and $\boldsymbol{v}(n) = [v(n),\ v(n-1),\ \ldots,\ v(n-K+1)]^{\mathrm{T}}$. The update rule of the FIR filter's weights in NLMS is given by

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + \mu \cdot s_{\mathrm{ANC}}(n)\boldsymbol{v}(n) \big/ \left(\xi + \|\boldsymbol{v}(n)\|_2^2\right) \tag{5}$$

where μ is the convergence factor, ξ is a small positive constant used to avoid division by zero, and $\|\cdot\|_2$ is ℓ_2-norm. The coarsely denoised auscultation signal $s_{\mathrm{ANC}}(n)$ provides a control signal and updates the filter's coefficients adaptively.

It has been shown in [26] that when the noise is additive and output of the adaptive filter $y(n)$ matches the ambient noise passed through an FIR system well, the ANC-NLMS error output $s_{\mathrm{ANC}}(n)$ is the optimal estimate of the target signal. NLMS also brings a fast convergence rate and low computational complexity. However, in real applications, a simple FIR filter can not sufficiently model the unknown system. Even if this unknown system is linear, we still do not know the number of orders of the FIR filter. Besides, the signal mixture in noisy auscultation sound generation is not guaranteed to follow an additive way. Hence, it is suspected that some residual noise ingredients may still exist in $s_{\mathrm{ANC}}(n)$, while some useful cardiopulmonary sound components may be carelessly subtracted from $d(n)$. On this account, a second stage is essential for the refinement of auscultation denoising.

2.2. The second denoising stage: DINC-Net

Conv-TasNet, a state-of-the-art audio/speech separation method, has been applied in tasks such as speaker extraction [27], echo suppression [28], and speech recognition [29]. The Conv-TasNet is a monaural source separation DNN, consisting of one encoder, one separation module, and one decoder. By using 1-D convolution, the encoder transforms the raw mixture waveform to a high-dimensional feature map, which is further multiplied by the mask generated from the separation module. The masked feature map is ultimately decoded to give separated sources in the decoder using 1-D transposed convolution. Inspired by the monaural Conv-TasNet, our proposed DINC-Net system consists of two encoders, one interactive denoising module, and one decoder, to consider the second stage denoising refinement problem. $y(n)$ and $s_{\mathrm{ANC}}(n)$ make up the inputs to the DINC-Net, and the final output $\widehat{s}(n)$ is the expected denoised auscultation sound result. The block diagram of the proposed DINC-Net is shown in Fig. 2. Compared to the Conv-TasNet, our DINC-Net has two main contributions. The first contribution is that two encoders are used to extract the features of $y(n)$ and $s_{\mathrm{ANC}}(n)$, and a deep encoder/decoder based on a stack of small-kernel filters with nonlinear activation functions is used to replace the original Conv-TasNet encoder/decoder using 1-D convolution. The second contribution is that we build interaction blocks from different feature dimensions to exchange information among different branches. Such an interaction scheme is specifically proposed for the two-channel setup, and will be proved to suppress the residual noise in the meanwhile of enhancing the useful signal components. Specific details about the DINC-Net are described as follows.

2.2.1. Deep encoder/decoder

The deep encoder built in our DINC-Net employs multiple convolutional layers to transform each frame of the waveform to effective implicit representations, and the deep decoder stacks multiple transposed convolutional layers to convert the implicit representations back to the desired waveform.

We create two deep encoders all with I layers to extract features of $y(n)$ and $s_{\mathrm{ANC}}(n)$. The first layer of our deep encoder is similar to that in the original Conv-TasNet encoder, implemented via a 1-D convolutional layer with N kernels to perform linear transformations of the input frame with L samples. A stack of 1-D dilated convolutional layers, with each layer having N kernels of size 3, follows the 1-D convolutional layer. The effctiveness of stacking dilated convolutional layers in building the encoder/decoder was verified in [30]. We use $I-1$ dilated convolutional layers with exponentially increasing dilation factors $1, 2, \ldots, 2^{I-2}$ for two deep encoders, and exponentially decreasing dilation factors $2^{I-2}, 2^{I-3}, \ldots, 1$ for the deep decoder. As what was pointed out in [31], dilated convolution has been used to increase the receptive field of convolutional layers and hence can efficiently extract features in temporal domain. A parametric rectified linear unit (PReLU) [32] is added to the output of each dilated convolutional layer, defined as

$$\mathrm{PReLU}(x) = \begin{cases} x, & \text{if } x \geqslant 0, \\ \alpha x, & \text{if } x < 0, \end{cases} \tag{6}$$

where α is a trainable scalar controlling the negative slope of the rectified activation unit. The outputs of the two encoders are implicit representations of $s_{\mathrm{ANC}}(n)$ and $y(n)$, denoted as $\boldsymbol{F}_{\mathrm{S,\,E}} \in \mathbb{R}^{T\times N}$ and $\boldsymbol{F}_{\mathrm{N,\,E}} \in \mathbb{R}^{T\times N}$, respectively, where T represents the feature length of decoder output, and N is the number of filters used in the encoder.

The deep decoder consists of four 1-D transposed dilated convolutional layers followed by one 1-D transposed convolutional layer, and the output is $\widehat{s} \in \mathbb{R}^{L\times 1}$. It is easy to derive the structure of the decoder, as it can be deemed as the mirror image of the encoder in the sense of recovering the implicit representations to the original size of input frame.

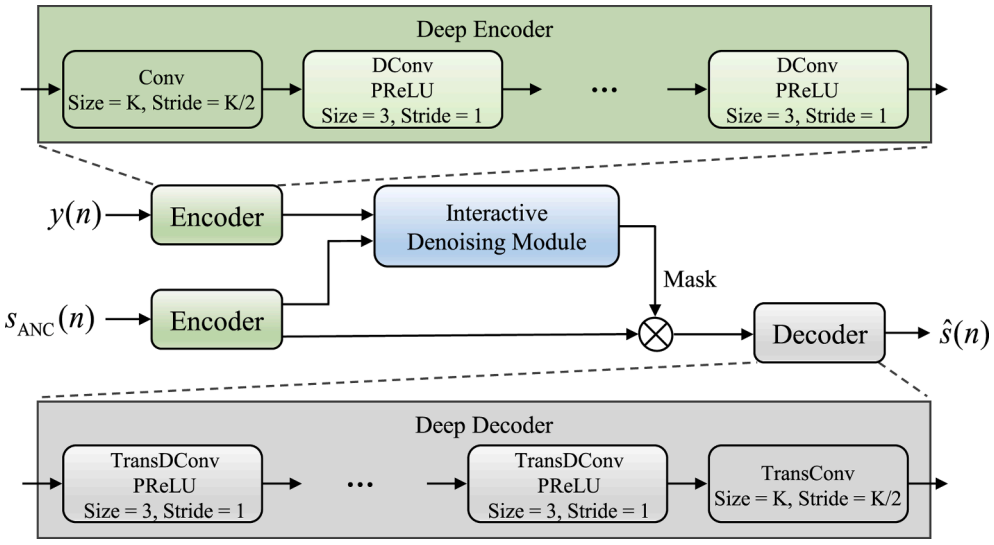


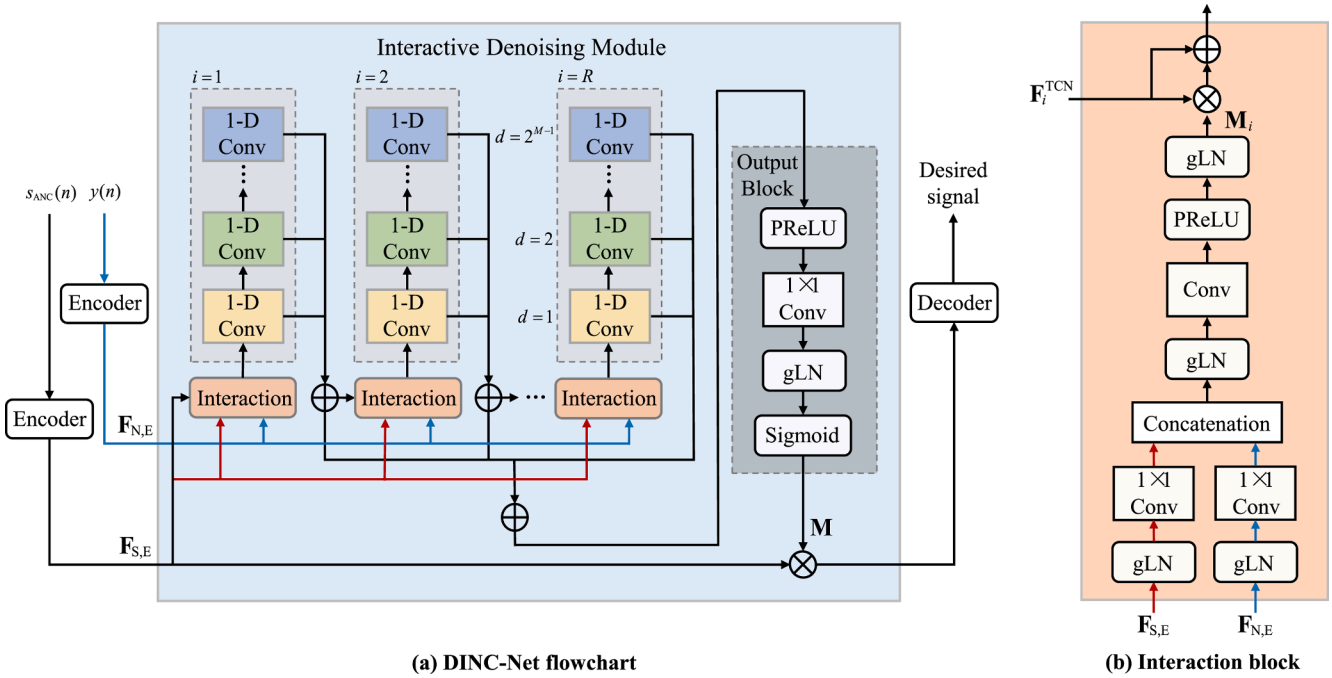
Fig. 2. The architecture of the proposed DINC-Net in the second denoising stage. Two deep encoders map two waveform inputs, i.e. coarsely denoised auscultation signal and estimated interference, to high-dimensional representations, and then an interactive denoising module calculates a mask for denoising refinement. A deep decoder reconstructs the denoised source waveform from the masked features. Conv and DConv represent 1-D convolutional layer and dilated convolutional layer, respectively. TransConv and TransDConv represent 1-D transposed convolutional layer and transposed dilated convolutional layer, respectively. PReLU is a nonlinear activation function formed by parametric rectified linear unit.

2.2.2. Interactive denoising module

This module possesses of three constituent parts, named temporal convolutional networks (TCNs), interaction blocks, and an output-end block. Fig. 3 (a) depicts the diagram of the proposed interactive denoising module and the role it plays in the second denoising stage.

There are R repeated TCNs in the interactive denoising module, and each TCN contains B 1-D Conv blocks whose dilation factors of their dilated convolutional layers are 1, 2, …, 2^{B-1}. The definition of 1-D Conv block is the same as that in Conv-TasNet. A 1-D Conv block has a 1×1 convolutional layer and a dilated convolutional layer, with each layer followed by PReLU and normalization. The output of each 1-D Conv block includes a residual path and a skip-connection path, where the residual path provides the input to the next block, and the skip-connection paths of all blocks in the i th TCN are summed up and serve as the output of this TCN, denoted as $\boldsymbol{F}_i^{\mathrm{TCN}}$. The input to 1-D Conv block is zero padded. By exploiting concatenated multi-scale feature maps generated by dilated convolutions with different dilation factors in parallel or cascade way, the TCN can achieve large receptive field size without increasing the size of kernels.

Through the first stage of denoising, in $s_{\mathrm{ANC}}(n)$ the residual noise may still exist, which is related to the noise components in $y(n)$. The interaction block is the key to extract cross-modality information from output feature maps of two encoders and feed them to the main branch. Fig. 3 (b) shows the design of interaction block. It has three input branches,



(a) DINC-Net flowchart **(b) Interaction block**

Fig. 3. Technical details about the proposed interactive denoising module in the second stage of denoising: (a) The flowchart of the interactive denoising module consisting of TCNs, interaction blocks, and an output-end block. The TCNs provide concatenated multi-scale feature maps generated by dilated convolutions with different dilation factors. Different colors in the 1-D convolutional blocks in TCNs indicate different dilation factors. The interaction block fuses the information of two encoders' output feature maps and TCN's outputs in the previous level. The aggregated features are processed to generate the final mask through output-end block; (b) The architecture of an interaction block. It has three input branches, including the output of the ith TCN for the main input branch and two auxiliary input branches given by encoders' outputs. Its output is fed into the following TCN.

including the output of the ith TCN $\boldsymbol{F}_i^{\mathrm{TCN}}$ for the main input branch and two auxiliary input branches given by $\boldsymbol{F}_{\mathrm{S,\,E}}$ and $\boldsymbol{F}_{\mathrm{N,\,E}}$. Firstly, the auxiliary branch inputs are normalized by global layer normalization (gLN) and then the number of channels is half reduced by 1×1 convolutional operations. Secondly, the two branches are merged by concatenating their channels, and the concatenated feature map further undergoes gLN, convolution layer, PReLU, and another gLN, to generate a multiplicative mask $\mathbf{M}_{\mathrm{i}}$ that predicts the cancellation and preservation areas of main branch. A gain representation $\boldsymbol{G}_i \triangleq \boldsymbol{F}_i^{\mathrm{TCN}} \odot \boldsymbol{M}_i$ is then obtained via element-wise multiplication. Finally, $\boldsymbol{F}_i^{\mathrm{TCN}}$ is added to $\boldsymbol{G}_i$ to obtain a "filtered" version of current feature map, to be fed into the next TCN. This process is given by

$$\boldsymbol{F}_{1,i} = \boldsymbol{F}_i^{\mathrm{TCN}} + \boldsymbol{F}_i^{\mathrm{TCN}} \odot \boldsymbol{M}_i, \quad i = 0,\ 1,\ \ldots,\ R-1 \tag{7}$$

It can be noticed that similar interaction blocks have shown their effectiveness in image processing [33] and speech enhancement tasks [34].

The last constituent part is the output-end block, which gives the output of a mask for denoising refinement. As displayed in Fig. 3 (a), the outputs of all TCNs are concatenated and fed to the output-end block. A PReLU is first used, and then the feature dimension is reduced by 1×1 convolutional operations, followed by gLN. A Sigmoid activation function is used to estimate the final mask $\boldsymbol{M} \in \mathbb{R}^{T \times N}$, with the same size as that of $\boldsymbol{F}_{\mathrm{S,\,E}}$ and $\boldsymbol{F}_{\mathrm{N,\,E}}$. The implicit representation of desired signal with denoising refinement $\boldsymbol{D}$ is calculated by applying the mask $\boldsymbol{M}$ to the implicit representation of $s_{ANC}(n)$ encoded as $\boldsymbol{F}_{\mathrm{S,\,E}}$, i.e.

$$\boldsymbol{D} = \boldsymbol{F}_{\mathrm{S,\,E}} \odot \boldsymbol{M} \tag{8}$$

where $\odot$ denotes element-wise multiplication. The deep decoder utilizes this implicit representation of $\boldsymbol{D}$ to reconstruct the waveform of final denoised cardiopulmonary sound signal.

2.2.3. Training objective

In training the proposed DINC-Net in the second denoising stage, the objective function is to minimize the negative scale-invariant source-to-noise ratio (SI-SNR) [35], which is defined as:

$$\text{SI-SNR} = 10\log_{10}\left(\left\|\boldsymbol{s}_{\text{target}}\right\|_2^2 \Big/ \left\|\boldsymbol{e}_{\text{noise}}\right\|_2^2\right) \tag{9}$$

where $\boldsymbol{s}_{\text{target}} = \langle \hat{\boldsymbol{s}}, \boldsymbol{s} \rangle \boldsymbol{s} / \|\boldsymbol{s}\|_2^2$, $\boldsymbol{e}_{\text{noise}} = \boldsymbol{s} - \boldsymbol{s}_{\text{target}}$, and $\hat{\boldsymbol{s}} \in \mathbb{R}^{L \times 1}$ and $\boldsymbol{s} \in \mathbb{R}^{L \times 1}$ are the estimated and ground truth cardiopulmonary sound signal vectors, respectively. Before engaged into loss function calculation, $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ are both normalized to ensure scale-invariance.

3. Experimental settings

It seems that there is a paradox when evaluating the performance of auscultation denoising methods, as we never know the ground truth of the clean cardiopulmonary sound signals. The lack of ground truth signal also hinders training of the proposed DINC-Net. In this paper, we evaluate the performance of the proposed denoising method via simulation experiments as well as real electronic stethoscope application results. In simulation experiments, we use synthetic noisy auscultation data generated by combining cardiac or respiratory sound signals with noises, extracted from public databases. We further integrate the denoising function using the trained model developed based on the simulation data, into an electronic stethoscope Mintti Smartho-D2 (Suzhou Melodicare Medical Technology Co., Ltd., China) [24], and evaluate the developed denoising prototype in realtime applications, including auscultation on healthy subjects and aortic stenosis patients.

3.1. Data and evaluation metrics for simulation

In simulation experiments, ground truth pure cardiopulmonary sound signals were used to generate synthetic noisy cardiopulmonary sound signals, to train the proposed DINC-Net, as well as to evaluate the denoising performances. The pure cardiopulmonary sound signals were extracted from two public data sources, where cardiac sound signals and respiratory sound signals were selected from the 2016 PhysioNet/CinC Challenge (PhysioNet) [36] database and the 2017 International Conference on Biomedical Health Informatics (ICBHI) [37] database, respectively. As in these public databases the cardiopulmonary sound recordings were collected in uncontrolled environments, how to pick out and identify some recordings as "pure" ones is the key issue. We used the following scheme for data screening. Firstly, the cardiac sound quality judgment algorithm developed in [38] and the respiratory sound quality judgment algorithm developed in [39] were performed on PhysioNet and ICBHI databases, respectively, to coarsely select cardiac, respiratory, or mixed cardiopulmonary sound recordings that are deemed as being of high quality. Secondly, these coarsely selected recordings were further reviewed by two auscultation experts, and only the recordings that are judged as noise-free by both experts were reserved. The judgment criterion was defined by subjective assessment: (1) No ambient noise can be heard; (2) The recorded cardiopulmonary sound contains cardiac sound, respiratory sound, or a mixture of them; (3) Data corruption by clipping distortion or friction is not present. Through this data screening procedure, 161 recordings from PhysioNet database and 90 recordings from ICBHI database were finally picked out and deemed as pure cardiopulmonary sound recordings.

We generated synthetic noisy cardiopulmonary sound signals by corrupting the pure signals with environmental sounds. The environmental sounds were provided by the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [40] and our own noise dataset named as "Hospital" (including 36 audio recordings, each one of 1-min length) collected in clinical environments. A synthetic noisy cardiopulmonary sound recording was given by summing the pure one with noise passed through an FIR system, given by

$$d(n) = s(n) + \sum_{m=0}^{M-1} h(m) v(n-m) \tag{10}$$

where $h(0),\ h(1),\ \ldots,\ h(M-1)$ are parameters of the FIR system, and M specifies the order. As in real applications the concrete form of the system to be identified is unknown, we used the following scheme to walk through probable scenarios: in generating each synthetic recording, M was randomly determined among 3–5, and the values of FIR parameters were randomly given from uniform distribution in $[-1, 1]$. Different SNR levels were covered.

A training dataset and a testing dataset were formed by utilizing the pure cardiopulmonary sound and noise recordings. It is worth noting that the pure cardiopulmonary sound recordings used in testing should differ from those used in training, to address individual variability and reflect reasonable generalization ability. A total of 141 PhysioNet recordings and 76 ICBHI recordings randomly selected out from the pure recordings were used for generating the training dataset, and the rest ones were used for the testing set. In forming the training dataset, 48 noise recordings of DEMAND in three environments (OOFFICE, OHALLWAY, and SPSQUARE) and our own 36 Hospital recordings were used for corrupting the cardiopulmonary sounds, with 5 randomly chosen SNR levels, from −5 dB to 15 dB with steps of 5 dB. The testing set includes 2 subsets: noise-contaminated PhysioNet recordings and ICBHI recordings. 5 SNR levels were considered, from −6 dB to 6 dB with steps of 3 dB, most of which are different from the implementation in building the training set. For each testing subset, two levels of denoising difficulty were considered, denoted by LEVEL-E and LEVEL-D, corresponding to easy and difficult noise cancellation tasks, respectively. In LEVEL-E, the utilized noise recordings had the same patterns as those used in the training set, while in LEVEL-D distinct patterns were used. When generating synthetic data, as the employed noise recordings were much longer, the noise recording was aligned with the

cardiopulmonary sound recording with a random starting point. The noise recordings were reused for 10 times in generating the training set. In simulation experiments, the input length of our proposed algorithm was fixed at 2 s, and the signal sampling rate was 8 kHz. Hence all the simulation data were resampled to 8 kHz and normalized, and further divided into 2 s segments, with 50 % overlapping. Table 1 briefly lists the specifications of the training/testing datasets.

Since there was no consistent standard for auscultation denoising performance evaluation, we chose two existing objective quality metrics in audio signal processing: the normalized covariance measure (NCM) [41] and the frequency-weighted segmental signal-to-noise ratio (fwSNRseg) [42]. To reflect the audibility of the denoised signal, the NCM calculates a weighted signal to noise quantity $\widehat{\mathrm{SNR}}_{\mathrm{NCM}}(s_p, \widehat{s}_p)$ at P bands, given by the normalized covariance of the spectral envelopes of s and $\widehat{s}$, i.e.

$$\mathrm{NCM}(s,\ \widehat{s}) = \frac{\sum_{p=1}^{P} c_p \times \widehat{\mathrm{SNR}}_{\mathrm{NCM}}(s_p, \widehat{s}_p)}{\sum_{p=1}^{P} c_p} \tag{11}$$

where c_p, $p = 1,\ 2,\ \ldots,\ P$ are band-importance weights. $P = 8$ bands whose center frequencies follow the Bark scale in [150, 4000] Hz were used, and the band-importance weights setting and the calculation of $\widehat{\mathrm{SNR}}_{\mathrm{NCM}}(s_p, \widehat{s}_p)$ followed the routine given by [43]. According to [44], the fwSNRseg exhibits the highest correlation with subjective signal quality assessment. It is essentially a weighted segmental SNR at critical frequency bands, defined by

$$\mathrm{fwSNRseg}(s,\ \widehat{s}) = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{j=1}^{J} w_j \mathrm{SNR}(j,\ m) \Bigg/ \sum_{j=1}^{J} w_j \tag{12}$$

where M is the total number of frames, J is the number of critical frequency bands, w_j is the weight at the jth band, $\mathrm{SNR}(j,\ m) \triangleq 10\log_{10}\left[|S(j,\ m)|^2/(|S(j,\ m)| - |\widehat{S}(j,\ m)|)^2\right]$, and $S(j,\ m)$ and $\widehat{S}(j,\ m)$ represent the spectral components of ground truth pure cardiopulmonary sound signal and denoised output of at the m_{th} frame and the jth critical frequency band, respectively. In this paper, the frame length was fixed at 30 ms, and $J = 25$ filters designed according to Articulation Index was used to specify the critical bands, whose weights followed the ones proposed in [45].

3.2. Data and evaluation metrics for realtime application

The ultimate goal of designing a practical noise cancellation algorithm for cardiopulmonary sound enhancement is to achieve a reasonable online performance in real auscultation applications, with the proposed noise cancellation method integrated and run in an electronic stethoscope. To consider the adaption to realtime application and complicated environments, some tiny adjustments compared to simulations were applied in building the online two-stage denoising prototype: the input data length was 0.5 s instead of 2 s, and all training sets and testing sets, truncated to segments with 0.5 s length, were employed to train the real implementation version of the proposed DINC-Net. To maintain the temporal continuity in the output of the developed online denoising prototype, cubic spline interpolation was used to bind the 0.5 s outputs. The developed online denoising prototype was deployed in Mintti Smartho-D2, a CE & FDA certificated electronic stethoscope.

Two real application scenarios were involved in performance evaluation. The first scenario considered auscultation on healthy subjects, with ambient speech interference. 10 healthy subjects, 22–25 years old, were recruited from Soochow University. 5 conventional cardiac auscultation positions were considered and each recording lasted for 10–30 s, with sampling rate = 8 kHz. During cardiac sound recording, a boy and a girl was talking aloud aside. The second scenario checked the auscultation enhancement performance on aortic stenosis patients in noisy clinical environments. 33 aortic stenosis patients, including 3 aortic stenosis levels (mild, medium, and severe) from 3 hospitals, participated in the experiments. All subjects or patients gave their signed informed consents before experiment. This study was approved by Ethics Committee of Soochow University (Approval No. SUDA20210923H02).

In order to facilitate performance evaluation, in the development mode of the deployed denoising prototype, the recorded data not only provided the denoised result but also included the raw noisy data and the ambient noise collected by the two microphones. The performance evaluation was performed segment-wise, and the data before or after denoising were divided into 2 s segments. We finally obtained 567 segments from normal recordings and 1140 segments from aortic stenosis patient recordings.

Due to the lack of ground truth pure cardiopulmonary sound, it is hard to use any objective metrics designed for simulation experiments to evaluate noise cancellation performance in real auscultation applications. In this paper, we built a discriminator, introduced from generative adversarial networks (GAN) [46], to automatically judge a denoised segment in real application as “acceptable” or “unacceptable”. The architecture of the built discriminator is presented in Table 2. We were interested in inspecting how many “unacceptable” segments can be turned into “acceptable” by the proposed two-stage noise cancellation algorithm. To avoid over-optimistic assessment, a conservative discriminator was trained: pure cardiopulmonary sound segments used in simulation experiments formed the “acceptable” class, while all the synthetic noisy ones formed the “unacceptable” class in the discriminator training set.

3.3. Implementation details and training setup

In the simulations as well as the real applications, the implementations of the proposed two-stage noise cancellation approach were similar, where the only difference was the input size. A well-established

Table 2
Detailed Parameter Setup for Discriminator.

Layer	Input Size	Channel	Kernel size	Stride	Dilation
Conv	1 * 16,384	8	8	4	–
DConv	8 * 4096	8	5	1	1
DConv	8 * 4093	8	5	2	2
DConv	8 * 2045	8	5	1	4
DConv	8 * 2037	8	5	2	8
Flatten	8 * 1011	–	–	–	–
Dense	8088	1024	–	–	–
Dense	1024	2	–	–	–
Softmax	2	–	–	–	–

Table 1
Specifications of Training Dataset and Two Testing Datasets with Different Levels of Denoising Difficulty.

Dataset		Noise Recordings	Noise Patterns	SNR (dB)	Segments (2 s)
Training dataset		84	{[OOFFICE, OHALLWAY, SPSQUARE], Hospital}	[−5, 0, 5, 10, 15]	14,823
Testing datasets	LEVEL-E	84	{[OOFFICE, OHALLWAY, SPSQUARE], Hospital}	[−6, −3, 0, 3, 6]	PhysioNet: 600 × 5 ICBHI: 756×5
	LEVEL-D	48	[STRAFFIC, TBUS, NFIELD]	[−6, −3, 0, 3, 6]	PhysioNet: 276 × 5 ICBHI: 504 × 5

form of ANC-NLMS in the first stage was employed, while the advantage of the proposed DINC-Net would be confirmed via ablation experiments. The issue how to choose a reasonable order of ANC-NLMS in the first stage would be addressed in Section 5. We also recognized that a small step size of NLMS helped improving the performance of the followed DINC-Net performance via simulation experiments. Hence, the order and the step size of ANC-NLMS in the first stage were set to 4 and 0.001, respectively. In the implementation of the proposed DINC-Net in the second stage, the tradeoff between denoising performance and model size was considered. In each encoder/decoder employed in the DINC-Net, the number of convolutional layers was $I = 5$, the number of channels in each convolutional layer was $N = 256$, and the kernel size for the first convolution layer in the encoder or the last transposed convolution layer in the decoder was $K = 16$. In the interactive denoising module, $R = 4$ repeated TCNs were used, and each TCN contained $B = 8$ 1-D Conv blocks. In each 1-D Conv block, the kernel size was 3, and the numbers of channels, including those in the residual paths and the skip-connection paths, was 256. In the interaction blocks, the kernel size for the convolutional layer was 16.

The training of the DINC-Net was implemented based on the PyTorch platform on a desktop equipped with two NVIDIA GeForce GTX 1080Ti GPUs. The Adam optimizer [47] with weight decay of 10^{-5} was used, and the batch size was 8. The learning rate was initially set to 10^{-3} and decreased by multiplying with 0.1 once the validation loss was not improved in three consecutive epochs, and the maximum number of epochs was 100. Specifically, if the validation loss is not improved in 6 consecutive epochs, the training would be terminated early. The trained model and denoising example are shared at: https://github.com/1406429350/DINC-Net.

4. Results

4.1. Simulation results

In this section, a full-scale validation of the advantages of the proposed two-stage noise cancellation approach via experiments on simulation dataset is displayed. First, ablation experiments will be carried out to show the necessity and effectiveness of each constituent part of DINC-Net. Second, quantitative evaluations of noise cancellation performance compared to the existing algorithms will be further performed at different categories of auscultation sound signals or different levels of denoising difficulties.

Ablation experiments were performed to justify the effectiveness of deep encoder/decoder and interaction block in the proposed DINC-Net. To verify that both deep encoder/decoder and interaction block contribute to denoising process, comparisons among several noise cancellation algorithms were performed. As in two-stage algorithms, the ANC-NLMS plays the role of coarse filtering followed by the DNN-based signal fine-tuning, the simple ANC-NLMS serves as the baseline algorithm and also the first denoising stage for 3 compared two-stage algorithms. Hence, the compared algorithms include: 1) only ANC-NLMS; 2) passing the single output of ANC-NLMS $s_{\mathrm{ANC}}(n)$ to a Conv-TasNet; 3) passing the two outputs $y(n)$ and $s_{\mathrm{ANC}}(n)$ to DINC-Net without deep encoder/decoder; 4) passing the single output $s_{\mathrm{ANC}}(n)$ to DINC-Net without interaction block. The LEVEL-E datasets including all SNRs and categories of auscultation sound signals were involved to calculate the evaluation metrics. The NCM and the fwSNRseg averaged on LEVEL-E datasets for the proposed algorithm as well as the 4 compared ones are summarized in Table 3, where the model sizes are also provided. It can be observed that the proposed algorithm achieved the highest values of averaged NCM and averaged fwSNRseg, and all two-stage algorithms showed superior denoising performances compared to only ANC–NLMS. For our proposed two-stage denoising algorithm, the vanishing of deep encoder/decoder or interaction block lead to performance degradation of DINC-Net, where it seems that the interaction block played a more important role in DINC-Net's outperformance. The advantages achieved by both the deep encoder/decoder and the interaction block are in line with expectations. The deep encoders empower the network to acquire larger temporal receptive field, hence extracting more features embeded in two channels of inputs. The interaction blocks take full account of the combination of dual-channel features and TCN outputs in various levels, and reserve and exploit all the information that can be used. It can also be noticed that even if deep encoder/decoder or interaction block were removed from DINC-Net, the dual-inputs setup still outperformed the single-input setup for the DNN in the second stage. The above ablation experiments prove the rationality of the architecture of the proposed two-stage noise cancellation algorithm, from the perspective of effect of the second denoising stage as well as its deep encoder/decoder and interaction block. It is worth mentioning that, the proposed algorithm suffered from an extra 32 % increment in model size compared to the two stage approach involving Conv-TasNet, while it achieved huge promotions in averaged NCM and fwSNRseg.

Table 3
Results of Ablation Experiments on LEVEL-E.

Method	Model size	NCM	fwSNRseg (dB)
No process	–	0.126	−5.662
ANC-NLMS (Only)	–	0.138	−5.567
NLMS + Conv-TasNet	**3.27M**	0.233	−1.037
Proposed	4.32M	**0.630**	**8.455**
Without deep encoder/decoder	3.93M	0.505	5.354
Without interaction block	3.66M	0.301	2.623

Performance comparisons in terms of evaluation metrics considering different levels of denoising difficulty, different databases, and different SNRs, have been performed. As aforementioned, the pure cardiopulmonary sound recordings were selected out from two public databases: PhysioNet and ICBHI databases, which were established to provide recordings of cardiac sounds and respiratory sounds, respectively. In fact, most of the existing auscultation denoising algorithms aimed at extracting either cardiac sounds or respiratory sounds out from noisy auscultation data. Hence, performance evaluations on testing datasets using synthetic data generated from PhysioNet and ICBHI databases would be displayed separately, and for each subset common algorithms as well as specific algorithms were performed for comparison. The common algorithms included one-stage ANC-NLMS and two-stage ANC + Conv-TasNet [21]. For the subset corresponding to PhysioNet, ANC-RLS [12], specifically designed for cardiac sound denoising, was applied for comparison. Multiband spectral subtraction (MBSS) [13], proposed for respiratory sound denoising, was compared in ICBHI-related subset.

Noise cancellation performances evaluated using NCM and fwSNRseg on the testing subset corresponding to easy denoising tasks (LEVEL-E) are displayed in Table 4. The noises contaminating cardiopulmonary sound signals in the training set and the testing set come from the same source of environmental noises. Specifically, we let most of the recordings in the testing set have different SNRs from those in the training set, which imitates a real denoising task where the unknown SNR can vary in a wide range. The fwSNRseg and the NCM were calculated on a per segment basis and then averaged over all segments in each level of SNR.

It can be observed from Table 4 that, in LEVEL-E the proposed noise cancellation algorithm achieved the highest NCM and fwSNRseg in each SNR level, for both cardiac sound and respiratory sound denoising tasks. Compared to the denoising results of one-stage NLMS, in the testing set generated from PhysioNet database our proposed algorithm achieved 0.572 increments in the NCM and 14.878 dB increments in the fwSNRseg, averaged in all SNR levels. For the ICBHI-related testing set, we got 0.412 and 13.165 dB for increments in the averaged NCM and the averaged fwSNRseg. The proposed algorithm did not exhibit much performance difference in denoising cardiac sounds and respiratory

Table 4
Comparison of Noise Cancellation Performances on LEVEL-E.

	Metrics	NCM					fwSNRseg (dB)				
	SNR (dB)	−6	−3	0	3	6	−6	−3	0	3	6
PhysioNet	No process	0.039	0.046	0.058	0.076	0.101	−7.896	−7.778	−7.586	−7.294	−6.776
	ANC-NLMS	0.045	0.055	0.071	0.093	0.123	−8.118	−7.926	−7.633	−7.187	−6.518
	ANC-RLS	0.048	0.060	0.077	0.102	0.135	−8.046	−7.836	−7.514	−7.023	−6.286
	NLMS + Conv-TasNet	0.134	0.166	0.207	0.260	0.314	−4.723	−3.775	−2.628	−1.270	0.150
	NLMS + DINC-Net (Proposed)	**0.583**	**0.622**	**0.655**	**0.682**	**0.707**	**5.057**	**6.223**	**7.403**	**8.584**	**9.742**
ICBHI	No process	0.132	0.156	0.184	0.217	0.256	−4.994	−4.621	−4.079	−3.315	−2.279
	ANC-NLMS	0.138	0.163	0.193	0.228	0.268	−5.009	−4.551	−3.898	−3.000	−1.829
	MBSS	0.246	0.277	0.308	0.344	0.385	−4.936	−4.460	−3.781	−2.856	−1.644
	NLMS + Conv-TasNet	0.146	0.186	0.239	0.305	0.375	−2.229	−1.223	0.090	1.733	3.502
	NLMS + DINC-Net (Proposed)	**0.533**	**0.572**	**0.614**	**0.648**	**0.681**	**7.384**	**8.536**	**9.618**	**10.581**	**11.421**

sounds, hence proving its strong adaptability. At all SNR levels involved in these simulation experiments, the proposed algorithm maintained robust denoising performances. It is also noticed that ANC + Conv-TasNet was not guaranteed to outperform an arbitrary-one-stage noise cancellation algorithm, indicating that the design of DNN in the second stage plays a crucial role in the noise cancellation task. Compared to single-channel denoising methods such as the Conv-TasNet, the proposed two-stage approach is two-microphone setup oriented. In the second stage, if only single-channel denoising models are used to deal with the coarsely denoised auscultation signal, the final performance critically relies on the model obtained by a limited size of training set and the useful information reserved in the coarsely denoised signal. In the proposed DINC-Net, the estimated interference given by the first stage is also used, and interaction blocks are designed to fuse all features that can be used. For this reason, the proposed method is expected to dynamically adapt to changes in ambient noise, and the experimental results verified its effectiveness.

Table 5 shows the averaged NCM and fwSNRseg calculated when the noise recordings used in the testing set had distinct patterns from those in the training set, i.e. LEVEL-D testing set was used for performance evaluation. It can be found that, compared to ANC-NLMS, the proposed two-stage denoising algorithm still achieved significant increments in the NCM and the fwSNRseg: (0.474, 13.511 dB) and (0.409, 11.430 dB) for PhysioNet and ICBHI related testing set, respectively. The increments of evaluation metrics in LEVEL-D showed a slight decline compared to those in LEVEL-E, which is natural as the denoising difficulty increased. Nonetheless, the proposed algorithms still yielded substantial outperformance on denoising both cardiac sounds and respiratory sounds, compared to other algorithms. Such an advantage may also be attributed to the dual-input setup and DINC-Net's interaction blocks. Different from the monaural denoising tasks, in the denoising problem addressed in this paper, information of the ambient noise provided by the auxiliary microphone can be fully utilized by the proposed method, even though the noise pattern was not considered in the training set.

4.2. Real application results

We deployed an online noise cancellation prototype in the electronic stethoscope Mintti Smartho-D2. This noise cancellation prototype implemented end-to-end noise cancellation for each 0.5 s segment and gave consistent output with interpolation between adjacent denoised segments. In the training process, the evaluation metrics calculated on validation set showed that the proposed two-stage algorithm yielded 0.452 and 13.230 dB increments in the averaged NCM and fwSNRseg compared to NLMS, respectively. Such a slight performance degrade compared to 2 s process is worth to suffer when the required input data length can be reduced substantially.

This denoising prototype was applied in auscultation on healthy subjects and aortic stenosis patients, and the performances were evaluated by our built discriminator, which judged the 2 s-segments divided from the recordings as "acceptable" or "unacceptable" for further analysis. Table 6 displays signal quality assessment results of the denoised output and the simultaneously recorded noisy data for the two auscultation scenarios. The corresponding acceptable rates of electronic stethoscope's output segments with and without denoising process is intuitively displayed in Fig. 4. It can be noticed that the proposed noise cancellation prototype greatly improved the auscultation signal quality compared to the unprocessed noisy data: in both of the two scenarios, the acceptable rates can be raised from low levels to not less than 95 %.

Table 6
Signal Quality Assessment by Discriminator in Real Auscultation Applications.

		Recorded segments	
		Unacceptable	Acceptable
Normal	No process	313	254
	Output of the denoising prototype	22	545
Abnormal (Aortic Stenosis)	No process	840	300
	Output of the denoising prototype	57	1083

Table 5
Comparison of Noise Cancellation Performances on LEVEL-D.

	Metrics	NCM					fwSNRseg (dB)				
	SNR (dB)	−6	−3	0	3	6	−6	−3	0	3	6
PhysioNet	No process	0.110	0.135	0.167	0.203	0.244	−6.014	−5.720	−5.283	−4.670	−3.863
	ANC-NLMS	0.124	0.152	0.185	0.222	0.263	−6.997	−6.525	−5.869	−5.019	−3.964
	ANC-RLS	0.129	0.157	0.191	0.228	0.268	−6.991	−6.514	−5.857	−4.989	−3.902
	NLMS + Conv-TasNet	0.255	0.293	0.332	0.376	0.415	−3.011	−2.237	−1.202	0.033	1.238
	NLMS + DINC-Net (Proposed)	**0.615**	**0.641**	**0.659**	**0.686**	**0.716**	**5.285**	**6.564**	**7.856**	**9.125**	**10.353**
ICBHI	No process	0.157	0.183	0.215	0.254	0.300	−3.758	−3.227	−2.493	−1.498	−0.194
	ANC-NLMS	0.178	0.206	0.240	0.280	0.326	−3.552	−2.727	−1.672	−0.379	1.179
	MBSS	0.319	0.350	0.391	0.434	0.469	−3.393	−2.744	−1.888	−0.696	0.970
	NLMS + Conv-TasNet	0.239	0.285	0.339	0.401	0.456	−0.062	0.900	2.030	3.305	4.614
	NLMS + DINC-Net (Proposed)	**0.591**	**0.625**	**0.661**	**0.689**	**0.711**	**8.078**	**9.077**	**10.061**	**10.982**	**11.800**

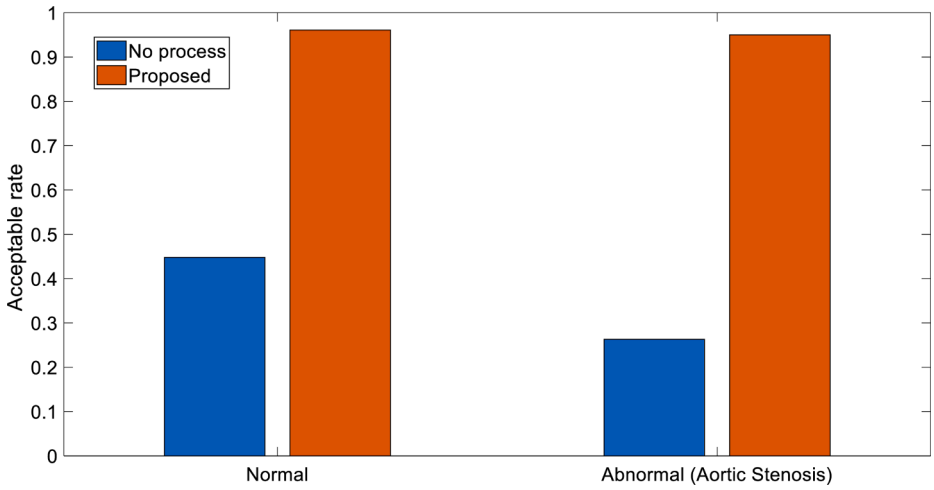


Fig. 4. The acceptable rates of electronic stethoscope's output segments before and after noise cancellation. The blue bars indicate acceptable rates without denoising by the proposed method, and the red bars indicate acceptable rates with the proposed denoising method involved. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Performance on healthy subjects was slightly better than that on aortic stenosis patients, as the level of difficulty increased in aortic stenosis auscultation denoising tasks: *e.g.* the aortic stenosis murmurs between S1 and S2 sound like some kinds of background noises. Anyway, the developed denoising prototype based on our proposed two-stage algorithm showed its effectiveness in normal and abnormal cardiac sound auscultation applications, even when judged by a "conservative" discriminator.

Fig. 5 and Fig. 6 show examples of noise cancellation results of auscultation on healthy subjects and aortic stenosis patients, respectively. Both of these two auscultation examples last for 12 s, and they are displayed in forms of time–frequency spectra. As aforementioned, in the development mode the raw noisy data and the ambient noise can also be recorded, and hence the compared algorithms can be carried out. The top of each figure displays the raw noisy auscultation data and the ambient noise, the middle rows show the denoising results of the one-stage ANC-NLMS and the two-stage NLMS + Conv-TasNet, and the output of our denoising prototype is plotted on the bottom of each figure. In order to facilitate visualization, only the spectrograms under 2 kHz for the denoised outputs are displayed. Both of these two examples illustrated the outperformance achieved by the two-stage algorithms compared to the one-stage algorithm, and our algorithm involving a dual-input interaction block in the second stage gave most favorable noise cancellation results. It is visualized from the example in Fig. 5 that, apart from noise cancellation, both the cardiac sounds and the respiratory sounds of a healthy subject were reserved. Fig. 6 shows that by our developed noise cancellation prototype, not only the noises lying on the areas without cardiac sounds were automatically removed, but also the

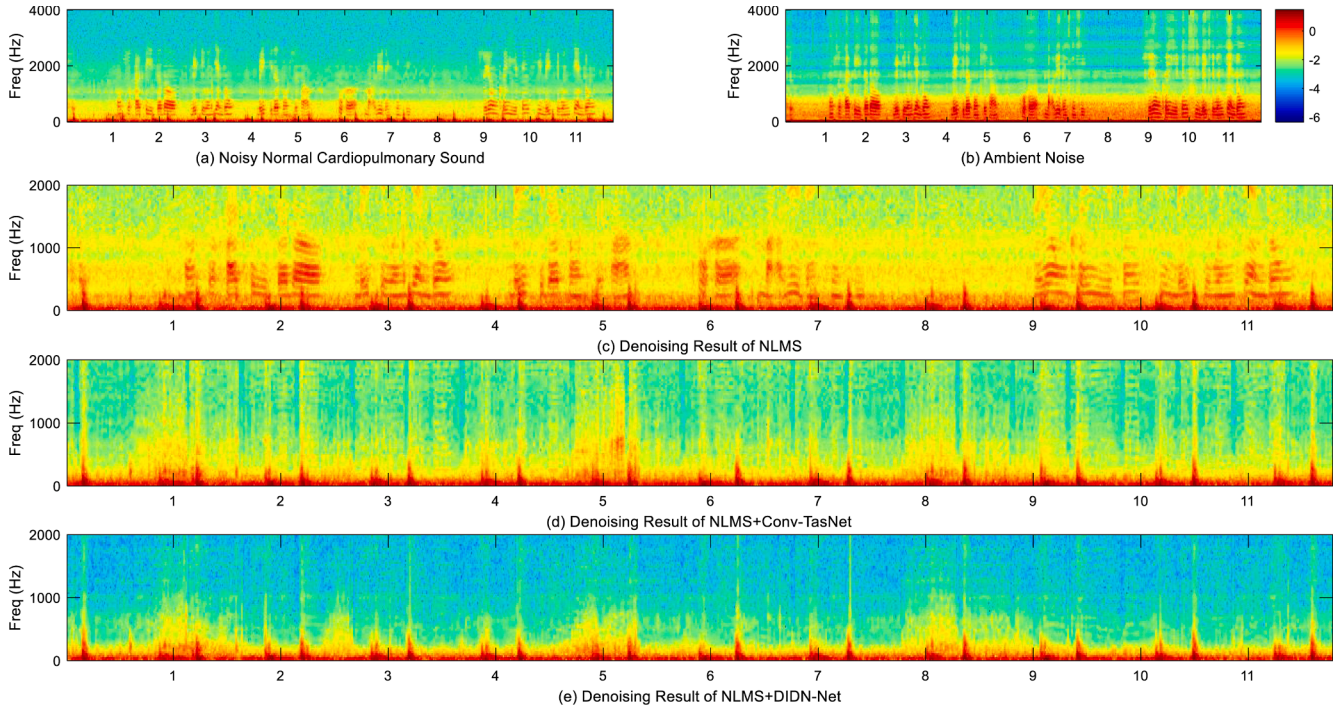


Fig. 5. A noise cancellation example for normal cardiopulmonary sound of a healthy subject. The signal quality was improved to various degrees by different methods: (a) NLMS; (b) NLMS + Conv-TasNet; (c) the proposed two-stage method NLMS + DIDN-Net. The most significant improvement in signal quality can be observed after denoising by the proposed method, where background speech interference and other noises were strongly suppressed.

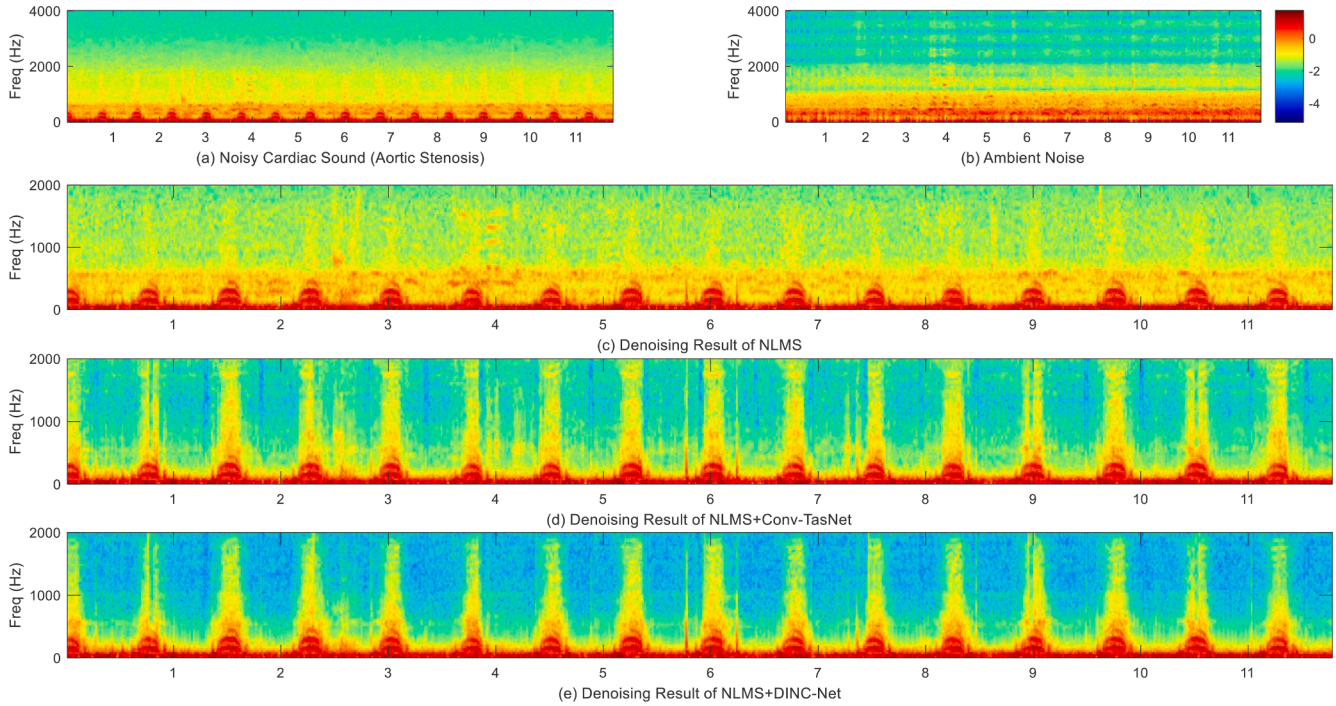


Fig. 6. A noise cancellation example for abnormal cardiac sound of an aortic stenosis patient. The proposed method not only eliminated speech interference and background noise, but also reserved the pathological features of cardiac sound for an aortic stenosis patient.

noise cancellation functioned when the noises were superimposed on cardiac sounds, including S1, S2, and murmurs.

5. Discussion

For real applications of the stethoscope in cardiopulmonary auscultation, noise cancellation is of great importance for collecting and further analyzing cardiopulmonary sound signals. Compared to the conventional stethoscopes, the electronic stethoscope has the advantage of employing various noise cancellation algorithms. Though monaural denoising methods were used to address this issue, the lack of "evidence" of ambient noises may lead to false cancellation of useful components, especially for auscultation on patients in complicated clinical environments. In the two-microphone setup, the reference channel collecting ambient noises right provides such "evidence", so our proposed two-stage noise cancellation approach is developed for the electronic stethoscope with an extra auxiliary microphone.

Our two-stage method contains a cascade of conventional ANC and DNN, realized by NLMS and our proposed DINC-Net. In fact, involving two-stage approach or ANN in this field has attracted some attention in very recent years. In [22], the two-stage ALE + ANN was designed for auscultation denoising, where the input to the simple ANN with only one hidden layer is the one-channel output of the ALE. In [48], ANNs with one or two hidden layers were combined with discrete wavelet transform for lung sound denoising, which demonstrated that ANN has the potential of auscultation enhancement. The key contribution of our work is that in the second denoising stage DNN, not only the coarsely denoised auscultation signal $s_{\mathrm{ANC}}(n)$ but also the estimated interference $y(n)$ from the first ANC stage were used as inputs, and a dual-channel interactive denoising module was designed to exploit and fuse the information in these two inputs for refined noise cancellation. To illustrate the advantage brought by the above innovation, a two-stage denoising method with single input to DNN in the second stage was carried out for comparison. To achieve the performance of this compared methodology to the maximum, Conv-TasNet, a state-of-the-art end-to-end denoising DNN, was used instead of a simple shallow neural network in the second stage. In our proposed DINC-Net, deep encoder/decoder and interactive denoising module are established. The deep encoder/decoder uses expansion factors to increase the temporal receptive field, and stack encoding/decoding layers hierarchically to transform the input waveform into a nonlinear latent space or the reverse. The interactive denoising module receives the encoded two-channel feature input and uses a masking scheme for auscultation denoising refinement. The advantages achieved by the contributions in the DINC-Net have been well displayed via ablation experiments. In addition, the experiments on simulation data as well as real applications showed the benefits by the proposed interactive denoising scheme. The results are as expected, verifying that both the residual noise ingredients in $S_{\mathrm{ANC}}(n)$ and the useful signal components in $y(n)$ can be efficiently addressed by the proposed method for denoising refinement.

In the first denoising stage, a traditional ANC-NLMS was used. To build the datasets for training our DINC-Net, the order of the unknown FIR system was randomly determined among $3 \sim 5$, and the corresponding parameters were also randomly given. In the performance evaluation experiments, the order of NLMS was fixed at 4, and the main concern was focused on the second denoising stage. In Table 7, the noise cancellation performances of our method involving various orders of ANC-NLMS evaluated at LEVEL-E sets were displayed. It can be observed that the proposed two-stage denoising method did not benefit from a high-order NLMS in the first stage, while on the contrary slight performance degrade would occur with the order growth. In real applications, we never know the complete information of unknown system and the way to contaminate cardiopulmonary sounds by noises. Furthermore, a complicated auscultation environment implies fast varying system and

Table 7
Denoising Performances in Variation of NLMS Filter Orders.

Evaluation metrics	Method	Filter order (N)				
		$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$
NCM	NLMS	0.138	0.139	0.147	0.145	0.145
	Proposed	0.630	0.597	0.571	0.557	0.523
fwSNRseg (dB)	NLMS	−5.567	−5.562	−5.401	−5.435	−5.468
	Proposed	8.455	8.005	6.302	6.195	5.814

unpredictable category of noises. Even so, the results in Table 7 suggest employing a simple form of ANC in the first stage for coarse noise cancellation and leaving the further denoising refinement to the designed DINC-Net in the second stage. The real experimental results using our developed noise cancellation prototype also gave some plausible support for this assumption.

The first limitation of the study in this paper stems from the signal quality issue of the used data: in fact we never know if a cardiopulmonary sound recording is truly without noise, which can only be subjectively assessed by auscultation experts or some artificial machine. Note that such a predicament also severely limited the application of the existing auscultation denoising methods. The second problem is that training DNNs requires a large amount of data as a driver. However, to form a complete auscultation database, the collection of cardiorespiratory sounds is a burdensome task and taking all the complicated acoustic environments into account is not easy. Finally, individual variability of cardiopulmonary sounds among patients remains an issue to be considered, although our auscultation enhancement method has been validated on aortic stenosis patients.

In future works, more cardiopulmonary sound signals recorded in controllable ambient noise environments would be added, and more categories of pathological cardiac sounds and respiratory sounds would be involved, to improve the robustness of the proposed noise cancellation method in real auscultation applications. In training the proposed denosing model, transfer learning can also be adopted. Starting from an initialised model trained for a large-scale audio/speech signal denoising task, transfer learning may further improve our model's performance. The feedbacks from physicians, who are users of noise-cancelling electronic stethoscopes implemented with our denoising prototype, will also be considered for further method improvement.

6. Conclusion

The noise cancellation problem in electronic stethoscope has been addressed in this paper. A two-stage noise cancellation approach was developed for cardiopulmonary sound denoising. The first denoising stage was implemented with a conventional ANC method for coarse noise cancellation, and the second stage was built by our proposed DINC-Net, whose dual-inputs included the coarsely denoised auscultation signal and the estimated interference from the first stage. The proposed DINC-Net used two deep encoders to extract features of dual-inputs, one interactive denoising module to exploit the mutual information of dual-inputs for denoising mask generation, and one deep decoder to give an ultimate denoised output. The noise cancellation performance on simulation experiments and real auscultation applications verified the advantages achieved by the proposed two-stage method, compared to the one-stage methods as well as the two-stage method without considering the interactive denoising module proposed in this paper. The proposed two-stage method provided a promising technical route for noise cancellation in real online auscultation applications of electronic stethoscopes. For future research, more categories of auscultation signals and physician users' feedbacks would be involved to improve the robustness of the proposed noise cancellation method. Transfer learning will also be introduced in the training of the proposed model to improve denoising performance.

CRediT authorship contribution statement

Chunjian Yang: Conceptualization, Methodology, Formal analysis, Software, Investigation, Visualization, Writing – original draft. **Neng Dai:** Conceptualization, Validation, Data curation, Visualization. **Zhi Wang:** Investigation, Conceptualization, Validation, Data curation. **Shengsheng Cai:** Validation, Investigation, Data curation. **Jiajun Wang:** Writing – review & editing, Resources, Project administration, Supervision. **Nan Hu:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by the Suzhou Science and Technology Project under Grant SYS2019029. The sponsors were not involved in study design, implementation or publication.

References

[1] M. Elhilali, J.E. West, The Stethoscope Gets Smart: Engineers from Johns Hopkins are giving the humble stethoscope an AI upgrade, IEEE Spectr. 56 (2) (2019) 36–41.
[2] I. McLane, D. Emmanouilidou, J.E. West, M. Elhilali, Design and comparative performance of a robust Lung auscultation system for noisy clinical settings, IEEE J. Biomed. Health. Inf. 25 (7) (2021) 2583–2594.
[3] S.R. Messer, J. Agzarian, D. Abbott, Optimal wavelet denoising for phonocardiograms, Microelectron. J. 32 (12) (2001) 931–941.
[4] O. Beya, B. Jalil, E. Fauvet, and O. Laligant, Empirical modal decomposition applied to cardiac signals analysis, in: Proc. of SPIE-IS&T Electronic Imaging, vol. 7535, Feb. 2010, pp. 1–11.
[5] M. K. Zia, B. Griffel, and J. L. Semmlow, Robust detection of background noise in phonocardiograms, in: Proc. first Middle East Conf. Biomed. Eng., 2011, pp. 130–133.
[6] A. Gavrovska, M. Slavkovic, I. Reljin, and B. Reljin, Application of wavelet and EMD-based denoising to phonocardiograms, in: Proc. Int. Symp. Signals, Circuits Syst., 2013, pp. 1–4.
[7] A. Mondal, I. Saxena, H. Tang, and P. Banerjee, A noise reduction technique based on nonlinear kernel function for heart sound analysis, IEEE J. Biomed. Health. Inf. 22(3) (2018) 775–784.
[8] N.S. Haider, Respiratory sound denoising using empirical mode decomposition, hurst analysis and spectral subtraction, Biomed. Signal Process. Control 64 (2021), 102313.
[9] L. Li, W. Xu, Q. Hong, F. Tong, and J. Wu, Classification between normal and adventitious lung sounds using deep neural network, in: Proc. 10th Int. Symp. Chin. Spoken Lang. Process., Oct. 2017, pp. 1–5.
[10] H.C. Shin, A.H. Sayed, W.J. Song, Variable step-size NLMS and affine projection algorithms, IEEE Signal Process. Lett. 11 (2) (2004) 132–135.
[11] N.Q. Al-Naggar, M.H. Al-Udyni, Performance of adaptive noise cancellation with normalized last-mean-square based on the signal-to-noise ratio of lung and heart sound separation, J. Healthcare Eng. (2018), 9732962.
[12] J. Gnitecki, Z. Moussavi, and H. Pasterkamp, Recursive least square adaptive noise cancellation filtering for heart sound in lung sounds recording, in: Proc. IEEE Eng. Med. Biol. Soc., 2003, pp. 2416–2419.
[13] D. Emmanouilidou, E.D. McCollum, D.E. Park, M. Elhilali, Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries, IEEE Trans. Biomed. Eng. 62 (9) (2015) 2279–2288.
[14] D. Gradolewski, G. Magenes, S. Johansson, W.J. Kulesza, A wavelet transform-based neural network denoising algorithm for mobile phonocardiography, Sensors 19 (4) (2019) 957.
[15] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F. Smolle-Juttner, H. Olschewski, and F. Pernkopf, Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks, in: Proc. EMBC, 2018, pp. 356–359.
[16] K.-H. Tsai, et al., Blind monaural source separation on heart and lung sounds based on periodic-coded deep autoencoder, IEEE J. Biomed. Health. Inf. 24 (11) (2020) 3203–3214.
[17] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, G.D. Clifford, Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration, IEEE Trans. Biomed. Eng. 68 (2) (2021) 650–663.
[18] S.Y. Lu, S.H. Wang, Y.D. Zhang, TBNet: a context-aware graph network for tuberculosis diagnosis, Comput. Methods Programs Biomed. 214 (2022), 106587.
[19] S.Y. Lu, S.H. Wang, Y.D. Zhang, Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm, Neural Comput. Appl. 33 (2021) 10799–10811.
[20] A. Li, W. Liu, C. Zheng, X. Li, Two heads are better than one: a two-stage complex spectral mapping approach for monaural speech enhancement, IEEE/ACM Trans. Audio. Speech, Lang. Process. 29 (2021) 1829–1843.
[21] X. Xiang, X. Zhang, H. Chen, Two-stage learning and fusion network with noise aware for time-domain monaural speech enhancement, IEEE Sig. Process. Lett. 28 (2021) 1754–1758.

[22] S. Rajkumar, K. Sathesh, N.K. Goyal, Neural network-based design and evaluation of performance metrics using adaptive line enhancer with adaptive algorithms for auscultation analysis, Neural Comput & Applic. 32 (2020) 15131–15153.
[23] Y. Luo, N. Mesgarani, Conv-TasNet: surpassing ideal time–frequency, magnitude masking for speech separation, IEEE/ACM Trans. Audio. Speech, Lang. Process. 27 (8) (2019) 1256–1266.
[24] Minttihealth: cardiopulmonary disease analysis and diagnosis system, Available from: <http://www.melodicare.cn/#/Product?productIndex=0>.
[25] R.M. Ramli, A.O.A. Noor, S.A. Samad, A review of adaptive line enhancers for noise cancellation, Austral. J. Basic Appl. Sci. 6 (6) (2012) 337–352.
[26] B. Widrow, et al., Adaptive noise cancelling: principles and applications, Proc. IEEE 63 (12) (1975) 1692–1716.
[27] C. Xu, W. Rao, E.S. Chng, H. Li, SpEx: Multi-scale time domain speaker extraction network, IEEE/ACM Trans. Audio. Speech, Lang. Process., Apr. 28 (2020) 1370–1384.
[28] H. Chen, T. Xiang, K. Chen, and J. Lu, Nonlinear residual echo suppression based on multi-stream Conv-TasNet, in: Proc. INTERSPEECH, 2020.
[29] J. Woo, M. Mimura, K. Yoshii, T. Kawahara, End-to-end music-mixed speech recognition, in: ProceedIngs of the Asia-Pacific Signal and Information ProcessIng Association Annual Summit and Conference, 2020, pp. 800–804.
[30] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, An empirical study of Conv-TasNet, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2020, pp. 7264–7268.
[31] A. Pandey and D. Wang, Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 6629–6633.
[32] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1026–1034.
[33] H. Wang, Z.-J. Zha, X. Chen, Z. Xiong, J. Luo, Dual path interaction network for video moment localization, in: Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 4116–4124.
[34] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, Y. Lu, Interactive speech and noise modeling for speech enhancement, Proc. AAAI 35 (2021) 14549–14557.
[35] Y. Luo and N. Mesgarani, TasNet: Time-domain audio separation network for real-time, single-channel speech separation, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 696–700.
[36] G.D. Cliffordet et al. “Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016, in: Proc. Comput. Cardiol. Conf., 2016, pp. 609–612.
[37] B. Rocha, D. Filos, L. Mendes, Vogiatzis et al., A respiratory sound database for the development of automated classification, in: Precision Medicine Powered by pHealth and Connected Health, 2018, pp. 33–37.
[38] H. Tang, M. Wang, Y. Hu, et al., Automated signal quality assessment for heart sound signal by novel features and evaluation in open public datasets, Biomed Res. Int. (2021) 1–15.
[39] A. Kala, A. Husain, E.D. McCollum, M. Elhilali, An objective measure of signal quality for pediatric lung auscultations, in: 2020 4second Annual International Conference of the IEEE Engineering, MedicIne & Biology Society, 2020, pp. 772–775.
[40] J. Thiemann, N. Ito, E. Vincent, The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings, J. Acoust. Soc. Amer. 133 (5) (2013) 3591.
[41] J. Ma, Y. Hu, P.C. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, J. Acoust. Soc. Amer. 125 (2009) 3387–3405.
[42] K. Kondo, Estimation of forced-selection word intelligibility by comparing objective distances between candidates, Appl. Acoust. 106 (2016) 113–121.
[43] Methods for calculation of the speech intelligibility index, ANSI-S3.5-1997-R2007, 1997.
[44] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation, IEEE/ACM Trans. Audio. Speech, Lang. Process 25 (4) (2017) 692–730.
[45] P.C. Loizou, Speech Enhancement: Theory and Practice, second ed., CRC Press, Boca Raton, FL, USA, 2013.
[46] S. Pascual, A. Bonafonte, J. Serrà, SEGAN: Speech enhancement generative adversarial network, in: Proc. Interspeech, 2017, pp. 3642–3646.
[47] D.P. Kingma and J.L. Ba, Adam: a method for stochastic optimization, in: Proc. Int. Conf. Learn. Represent., 2015, pp. 1–41.
[48] M.F. Pouyani, M. Vali, M.A. Ghasemi, Lung sound signal denoising using discrete wavelet transform and artificial neural network, Biomed. Signal Process. Control 72 (2022), 103329.