Biomedical Signal Processing and Control 49 (2019) 173–180

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Robust S1 and S2 heart sound recognition based on spectral restoration and multi-style training

Yu Tsao[a], Tzu-Hao Lin[a], Fei Chen[b], Yun-Fan Chang[c], Chui-Hsuan Cheng[c], Kun-Hsi Tsai[c,*]

[a] *Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei, Taiwan*
[b] *Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen, China*
[c] *iMediPlus Inc., Hsinchu, Taiwan*

ARTICLE INFO

Article history:
Received 28 January 2018
Received in revised form
15 September 2018
Accepted 22 October 2018

Keywords:
Deep learning
Spectral restoration
Multi-style training
Robust heart sound recognition
S1 and S2 recognition

ABSTRACT

Recently, we have proposed a deep learning based heart sound recognition framework, which can provide high recognition performance under clean testing conditions. However, the recognition performance can notably degrade when noise is present in the recording environments. This study investigates a spectral restoration algorithm to reduce noise components from heart sound signals to achieve robust S1 and S2 recognition in real-world scenarios. In addition to the spectral restoration algorithm, a multi-style training strategy is adopted to train a robust acoustic model, by incorporating acoustic observations from both original and restored heart sound signals. We term the proposed method as SRMT (spectral restoration and multi-style training). The experimental procedure in this study is described as follows: First, an electronic stethoscope was used to record actual heart sounds, and the noisy signals were artificially generated at different signal-to-noise-ratios (SNRs). Second, an acoustic model based on deep neural networks (DNNs) was trained using original heart sounds and heart sounds processed through spectral restoration. Third, the performance of the trained model was evaluated using the following metrics: accuracy, precision, recall, specificity, and F-measure. The results confirm the effectiveness of the proposed method for recognizing heart sounds in noisy environments. The recognition results of an acoustic model trained on SRMT outperform that trained on clean data with a 2.36% average accuracy improvement (from 85.44% and 87.80%), over clean, 20dB, 15dB, 10dB, 5dB, and 0dB SNR conditions; the improvements are more notable in low SNR conditions: the average accuracy improvement is 3.87% (from 82.83% to 86.70%) in the 0dB SNR condition.

1. Introduction

Auscultation serves as an effective and reliable method for performing automatic diagnosis of certain lethal cardiac disorders (e.g., valvular heart disease and congestive heart failure) at a low cost. In an automatic cardiac-disorder-diagnosis system, the recognition of the first heart sound (S1) and the second heart sound (S2) plays an essential role. Generally speaking, automatic heart sound recognition methods can be divided into unsupervised and supervised classification ones. An unsupervised classification method is usually derived based on the specific characteristics of S1 and S2 signals; notable approaches include the envelogram analysis [4] and the high frequency-based wavelet decomposition [5]. A supervised classification method usually involves a pattern recognition model, which learns the patterns of the two heart sound signals; successful models include neural network (NN) [6] and decision tree [7]. Meanwhile, some methods are derived based on the regularity of S1–S2 intervals with assumed average heart rates [8,9].

In our previous study [10], we have proposed a supervised method for S1 and S2 recognition that uses voiceprint analysis, deep learning, and Mel-frequency cepstral coefficients (MFCCs). MFCCs are features commonly used in acoustic event classification [11,12]. The Mel-frequency cepstrum is designed based on the frequency-domain perception system of a human ear. Moreover, we used a heart sound activity detector (HSAD), rather than manual means, to identify the segments of S1 and S2 heart sounds from an entire audio recording of heart sounds [13]. Subsequently, a *k*-means algorithm [14] was used to cluster a heart sound segment into two centroids, which were then used to form a long vector. Finally, S1 and S2 were classified using deep neural network (DNN) based classifiers. The experimental results show that the system can achieve high recognition performance in a clean recording

* Corresponding author.
E-mail address: peter.tsai@imediplus.com (K.-H. Tsai).

https://doi.org/10.1016/j.bspc.2018.10.014
1746-8094/© 2018 Published by Elsevier Ltd.

condition [10]. Despite the satisfactory recognition performance achieved in a clean condition, the performance robustness in a noisy condition is also an important consideration that determines the applicability of the system in real applications. In this study, we investigated the system robustness using noisy heart sound signals. To further improve the system robustness, we adopted a spectral restoration algorithm to reduce noise components from the noisy heart sound signals. Furthermore, a multi-style training strategy is used to estimate a robust acoustic model by incorporating diverse acoustic information from both original and restored acoustic heart sound signals.

A spectral restoration algorithm estimates a gain function for reducing noise in the frequency domain to retrieve cleaner sound spectra from noisy observations. Well-known spectral restoration methods include the Wiener filter [15] and spectral subtraction (SS) [16], with their related extensions [17,18]. Additionally, some spectral restoration methods are based on probabilistic models, such as the minimum mean square error (MMSE) [19–21], maximum a posteriori spectral amplitude (MAPA) [22,23], maximum likelihood spectral amplitude (MLSA) [24,25], and generalized MAPA (GMAPA) [26]. In [27], we investigated the integration of spectral restoration and deep-learning-based acoustic modeling on a robust speech recognition task. The experimental results showed that a direct combination of these two approaches may not effectively improve the speech recognition performance. A possible reason is that although spectral restoration can suppress noise components, the restored signals may suffer from signal distortions and still differ from clean signals. Motivated by the findings in [27], the present study proposes an alternative way to integrate spectral restoration and deep-learning-based acoustic modeling. Since the characteristics of DNNs can learn different patterns of sound signals [28–31], we use the spectral restoration method to generate useful and discriminative heart sounds as augmented data for training the acoustic model to improve performance robustness.

We evaluated the proposed approach using a noisy heart sound dataset, which was formed by artificially adding noise signals to clean heart sounds at varying SNR levels. The clean heart sounds were recorded by an electronic stethoscope in a quiet environment. The noise signals, which included electrical and background noises from the environment, were recorded using the same electronic stethoscope. The MMSE method and the minimum controlled recursive averaging (MCRA) noise tracking algorithm [32] were used to perform spectral restoration in this study. Four DNN-based acoustic models were prepared, based on the following different training sets: (1) clean heart sounds, (2) noisy heart sounds, (3) restored heart sounds, and (4) spctral restoration and multi-style training, namely a combination of clean, noisy, and restored heart sounds. The same noisy testing heart sounds were used to evaluate the performance of these four models. Experimental results showed that the model based on the multi-style training gave the best recognition accuracy under various noisy conditions, showing that the multi-style training approach can effectively improve the performance robustness over the conventional heart sound recognition system, as proposed in [10].

The rest of this paper is structured as follows. Section 2 describes the spectral restoration algorithm and multi-style training strategy. Section 3 introduces the S1 and S2 recognition system used in this study. Section 4 illustrates the experimental setup and results. Finally, Section 5 presents conclusions.

2. Spectral restoration and multi-style training

This section provides an overview of spectral restoration, which is used to reduce noise components in noisy heart sound signals. Assume that a pure signal $s[n]$ is affected by background noise $v[n]$ to produce a noisy signal $y[n]$. The goal of spectral restoration is to calculate restored signal $\hat{s}[n]$ from $y[n]$, as close to $s[n]$ as possible. Since the restored signal $\hat{s}[n]$ is less affected by noise, the recognition system should achieve better results by properly incorporating the information of $\hat{s}[n]$.

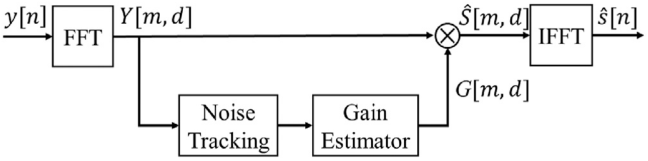

Fig. 1. Block diagram of a spectral restoration system.

2.1. Spectral restoration processing

Fig. 1 presents the overall procedure of the spectral restoration method, which consists of four parts: short-time Fourier transform (STFT), noise tracking, gain estimator, and inverse short time Fourier transform (ISTFT). In the time domain, the noisy signal can be represented as

$$y[n] = s[n] + v[n] \tag{1}$$

where n is the time index. The noisy signal $y[n]$ is first transformed into the frequency domain using the STFT. Then, we obtain:

$$Y[m,d] = S[m,d] + V[m,d],\ 0 \le d \le D-1, \tag{2}$$

where d is the frequency bin of ω_d with $\omega_d = 2\pi d/D$, m is the frame index, and $S[m,d]$ and $V[m,d]$ are the clean signal and noise spectra, respectively. For ease of notation, we denote $Y[m,d]$, $G[m,d]$, $S[m,d]$, and $V[m,d]$ as Y, G, S, and V, respectively, in the following discussion.

Next, when representing Y and S by their amplitude and phase components, we have:

$$Y = Y_q exp\left(j\theta y\right), \tag{3}$$

$$S = S_q exp\left(j\theta s\right), \tag{4}$$

where $Y_q = \left|Y\right|$, $S_q = \left|S\right|$, $\theta y = \angle Y$, and $\theta s = \angle S$. The a priori SNR ξ_q and the a posteriori SNR γ_q are defined as $\xi_q = \sigma_s^2/\sigma_v^2$ and $\gamma_q = Y_q^2/\sigma_v^2$, where $\sigma_s^2 = E\left[\left|S\right|^2\right]$, $\sigma_v^2 = E\left[\left|V\right|^2\right]$, and q denotes the amplitude part of the signals. Based on the a priori SNR ξ_q and a posteriori SNR γ_q, we compute a gain function G, which is used to filter Y to obtain the estimated $\hat{S}$. Generally, we adopt the phase of the noisy input as the enhanced sound by assuming:

$$exp\left(j\hat{\theta}s\right) = exp\left(j\theta y\right), \tag{5}$$

where $\hat{\theta}s$ denotes the estimated phase for the restored signal. Then, the restored signal is computed as

$$\hat{S} = \hat{S}_q exp\left(j\theta y\right) = G \cdot Y \tag{6}$$

Finally, the ISTFT is applied to generate the restored sound signal $\hat{s}[n]$.

2.2. MMSE based spectral restoration method

By assuming that both the clean sound and noise spectra can be modeled by Gaussian distributions, the conditional probability density function (PDF), $p\left(Y|S_q, \theta_S\right)$, can be derived as

$$p\left(Y|S_q, \theta_S\right) = \frac{1}{\pi\sigma_v^2} exp\left(\frac{-\left|V\right|^2}{\sigma_v^2}\right). \tag{7}$$

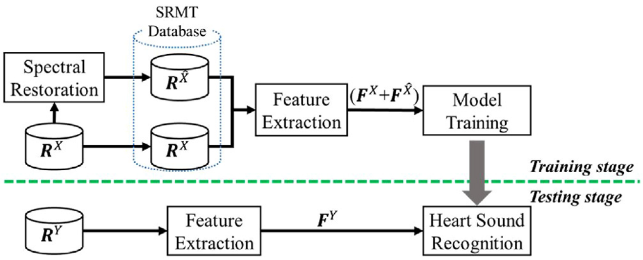

Fig. 2. Flowchart of the proposed SRMT strategy.

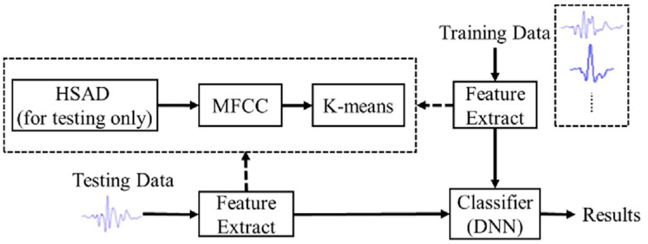

Fig. 3. Framework for identifying S_1 and S_2 system.

The amplitude and phase components of the complex Gaussian random variables with zero mean are known to be statistically independent [18]. Thus, $p\left(S_q, \theta_S\right)$ becomes

$$p\left(S_q, \theta_S\right) = p\left(S_q\right) \cdot p\left(\theta_S\right), \tag{8}$$

where $p\left(S_q\right)$ is modeled by the Rayleigh distribution

$$p\left(S_q\right) = \frac{2S_q}{\sigma_s^2} exp\left(\frac{-S_q^2}{\sigma_s^2}\right), \tag{9}$$

where σ_s^2 denotes the hyper-parameter in the density. $p\left(\theta_S\right)$ is modeled by a uniform density with

$$p\left(\theta_S\right) = \frac{1}{2\pi}, \tag{10}$$

The spectral amplitude of the MMSE estimator is given by the conditional expectation

$$\hat{S}_q = E\left[S_q|Y\right] = \int_0^\infty S_q p\left(S_q|Y\right) dS_q$$
$$= \frac{\int_0^\infty \int_0^\pi S_q p\left(Y|S_q, \theta_q\right) p\left(S_q, \theta_S\right) d\theta_S dS_q}{\int_0^\infty \int_0^\pi p\left(Y|S_q, \theta_S\right) p\left(S_q, \theta_S\right) d\theta_S dS_q}. \tag{11}$$

By substituting Eqs. (7) and (8) into Eq. (11), combined with some derivations, the MMSE-based gain function, G_{MMSE}, can be expressed as

$$G_{MMSE} = \Gamma\left(\frac{3}{2}\right)\frac{\sqrt{\delta}}{\gamma_q} exp\left(\frac{-\delta}{2}\right)\left[\left(1+\delta\right) I_0\left(\frac{\delta}{2}\right) + \delta I_1\left(\frac{\delta}{2}\right)\right], \tag{12}$$

where $\delta = \left[\xi_q / \left(1+\xi_q\right)\right]\gamma_q$; $\Gamma(\cdot)$, $I_0(\cdot)$, and $I_1(\cdot)$ denote the gamma function, zero-order modified Bessel function, and first-order modified Bessel function, respectively. The enhanced sound spectrum for the MMSE estimator can then be estimated by $\hat{S} = G_{MMSE} \cdot Y$.

2.3. Multi-style training criterion

The multi-style training strategy is a simple, yet effective approach to improving performance robustness of acoustic models. This approach estimates acoustic models by using sound data from diverse acoustic conditions (e.g., collecting sound data in different recording conditions or injecting noise into clean audio recordings) [33,34]. Several techniques have been implemented using the multi-style training approach, for instance: mixed sound data [35], noise injection [31], and mixed bandwidth data [36].

The main concept of spectral restoration with multi-style training, termed SRMT, in the following discussion is to improve the classification ability based on a collection of diverse training patterns (original and restored sound), to determine more accurate decision boundaries under heterogeneous conditions. In this study, the MMSE algorithm, as presented in the previous section, is used to generate restored sound data. Fig. 2 presents the architecture of the SRMT system, which consists of 1) spectral restoration to generate additional training data; 2) feature extraction from sound signals; 3) acoustic model training; and 4) a testing stage to perform recognition on the testing data using the trained acoustic model.

The procedure of the SRMT approach consists of training and testing stages, as can be noted in Fig. 2. In the training stage, the original training data $\boldsymbol{R}^X$ is processed to generate restored sound data $\boldsymbol{R}^{\hat{X}}$. Next, $\boldsymbol{R}^{\hat{X}}$ and $\boldsymbol{R}^X$ are augmented to form a larger training dataset. Then, the sound data $\boldsymbol{R}^X$ and $\boldsymbol{R}^{\hat{X}}$ are converted to the acoustic features $\boldsymbol{F}^X$ and $\boldsymbol{F}^{\hat{X}}$, respectively, which are then used to train the acoustic model. In the testing stage, feature extraction is carried out over the test sound data $\boldsymbol{R}^Y$ to generate the acoustic features $\boldsymbol{F}^Y$, which are finally used to test the recognition with the trained acoustic model.

3. DNN-based heart sound recognition system

Fig. 3 shows the framework of the DNN-based heart sound recognition proposed in our previous study [10]. First, S1 and S2 heart sound segments were converted into acoustic features. Second, the training acoustic features were used to train the DNN. Finally, the testing data was classified by the DNN classifier.

3.1. Feature extraction

The feature extraction process includes three parts: (1) HSAD, which preprocesses the entire sound signal and locates heart sound areas, (2) MFCC feature extraction, and (3) k-means clustering, which processes heart sound segments to yield final features.

3.1.1. Heart sound activity detector

The sound activity detection process is often applied in speech coding and acoustic event recognition. It is also used to preprocess classification problems to improve the recognition capability. The objective function of the HSAD algorithm is to identify acoustic segments as S1 or S2. Ideally, the heart sound signals received by a stethoscope are larger than noise. Therefore, the heart sound segments were determined according to the energy level of the audio segments.

3.1.2. MFCC feature extraction

MFCC feature extraction entails six steps: 1) pre-emphasis, 2) windowing, 3) fast Fourier transform (FFT), 4) Mel-filtering, 5) nonlinear transformation, and 6) discrete cosine transform (DCT). It has been widely shown that pattern recognition is enhanced when differential parameters are used to describe temporal characteristics. A differential cepstral parameter can be defined as the slope of a parameter with relation to time, which represents dynamic changes in cepstral parameters. Therefore, three times of dimensions of original features were obtained after appending velocity and acceleration features [37].

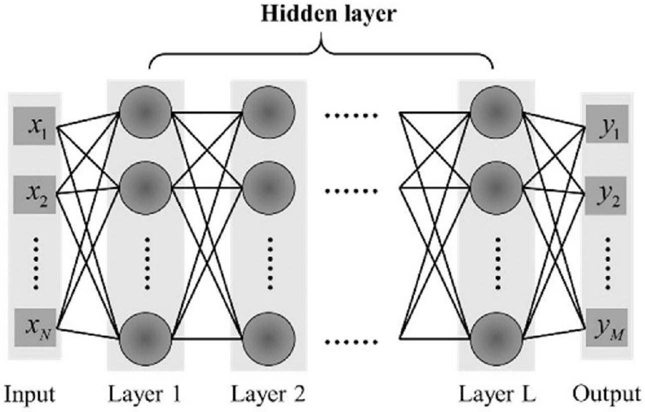

Fig. 4. The architecture of a DNN model.

3.1.3. K-means algorithm

The primary objective of the *k*-means algorithm is to combine numerous data points to form representative ones, which can be referred to as cluster centers. In this study, the number of cluster center was two, and the two groups of cluster centers formed by the MFCC of each heart sound segment were concatenated into a set of super-vectors as the final features. From our observation in most cases, one of the two clusters represents the features located in the central part, and the other contains the features in the outer part of the entire heart sound segment.

3.2. Deep neural network

An NN is a mathematical model that imitates the structure and functionality of a biological NN to perform classification or regression. Numerous NN models have been proposed to solve problems in different fields, such as the feed-forward network, the Hopfield network, and the radial basis function network. Moreover, a multilayer NN, commonly known as a DNN, has recently been demonstrated to be effective in a number of tasks, such as speech processing [38,39], speech recognition [40–42], and visual pattern recognition [43,44]. A DNN uses a plurality of hidden layers to strengthen its classification and regression capabilities. Fig. 4 illustrates the framework of a DNN model. The input ($\boldsymbol{a}_l$) and output ($\boldsymbol{a}_{l+1}$) of each hidden layer of the model are defined as follows:

$$\boldsymbol{a}_l = F\left(\boldsymbol{W}_l \boldsymbol{a}_{l-1} + \boldsymbol{b}_l\right), l = 1, 2, \ldots, L, \tag{13}$$

where $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are the weight matrix and bias of the l -th layer, respectively, $F(\cdot)$ is the activation function, and L is the number of hidden layers. On the basis of Eq. (13), the output of the current hidden layer is used as the input of the next. For the first layer, we have $\boldsymbol{a}_0 = \boldsymbol{x}$, where $\boldsymbol{x}$ is the input data. Moreover, the last output layer uses the softmax function, $C(\cdot)$, to yield the classification results, which can be expressed by

$$\hat{\boldsymbol{y}} = C\left(\boldsymbol{a}_L\right) \tag{14}$$

where $\hat{\boldsymbol{y}}$ is the classification result. The parameters of the DNN are estimated as follows:

$$\hat{\Lambda} = argmin_{\Lambda}\{Q\left(\boldsymbol{x}, \boldsymbol{y}; \Lambda\right) + \gamma R(\boldsymbol{W}) + \eta\rho(\boldsymbol{A})\} \tag{15}$$

where $\boldsymbol{y}$ is the correct label, $\Lambda = \left\{\boldsymbol{W}_l, \boldsymbol{b}_l, l = 1, 2, \ldots, L\right\}$ is the set of parameters of the DNN, $Q\left(\boldsymbol{x}, \boldsymbol{y}; \Lambda\right)$ is a loss function, $R(\boldsymbol{W})$ is a regularization function, and $\rho(\boldsymbol{A})$ is a sparsity penalty; the latter two are used at the DNN training stage to mitigate overfitting. Backpropagation is conducted to estimate the parameter set, Λ, and the estimation process is detailed in [10].

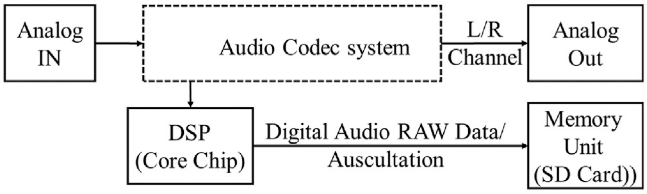

Fig. 5. Block diagram of signal processing in the electronic stethoscope (IMEDIPLUS DS 301) used for data collection.

4. Experiments

The feature extraction process conducted in this experiment is described in Section 3. After MFCCs were extracted, the *k*-means algorithm was used to obtain super-vectors as final features. At the training stage, heart sounds were manually segmented and labeled either S1 or S2 by a medical doctor. We used clean heart sounds with noise at different SNRs, as well as heart sounds that were processed through spectral restoration to train the DNN model. During the testing stage, the trained DNN model was used to recognize the heart sounds with noise at different SNRs.

4.1. Experimental setup

The data collection procedure, which entailed recording the normal sounds of the human heart and typical environmental and mechanical sounds, was approved by a local institutional review board. Before the trial, two qualified and licensed physicians were sought to define the correct locations and method of heart sound measurement. Subjects were asked to wear a thin cloth. The material of the clothes was not particularly limited. The heart sounds were recorded with a handheld electronic stethoscope (IMEDIPLUS DS301 [45]). The recording condition was clean, and each location was recorded twice, with 10 s each time. The recorded data were checked again by another group of physicians to ensure the recording quality before conducting experiment. Fig. 5 depicts the block diagram of the signal processing in the electronic stethoscope adopted to collect data. In Fig. 5, the Analog In unit receives sounds, the Audio Codec system contains the A/D, filter switching, and D/A units. The A/D (analog-to-digital) unit transforms an analog signal into a digital signal at a sampling frequency of 48 kHz, the filter switching unit emphasizes signals in lower frequency bands, and the D/A (digital-to-analog) unit reconstructs the digital signal into an analog form. The digital signal processing (DSP) unit processes the digital signal for storage in the memory unit, and the memory unit stores the signal for further analysis.

We collected two sets of data: a training set and a testing set. The training set comprised 313 S1 heart sounds and 313 S2 heart sounds, which were collected from 11 male and five female subjects. Processed by the HSAD, the testing set comprised 87 S1 heart sounds and 87 S2 heart sounds, which were collected from three male and three female subjects. In this study, we assume that the HSAD can perfectly detect heart sounds even at low SNR levels.

Fig. 6 depicts the locations of heart sound recordings, which are the pulmonary valve auscultation area ② and the second aortic valve auscultation area ③. We prepared noisy training and testing sets by artificially contaminating the clean training and testing heart sounds. The noise signals included electrical noise and background noise from the environment, such as human speech and fricative noises. These noise sources were recorded using the same electronic stethoscope shown in Fig. 5 and then divided into two parts, one for generating training data and the other for generating testing data. Therefore, the noise components in the training and testing sets were similar, but not completely the same. We considered this setup similar to the real-world scenarios, where the noise

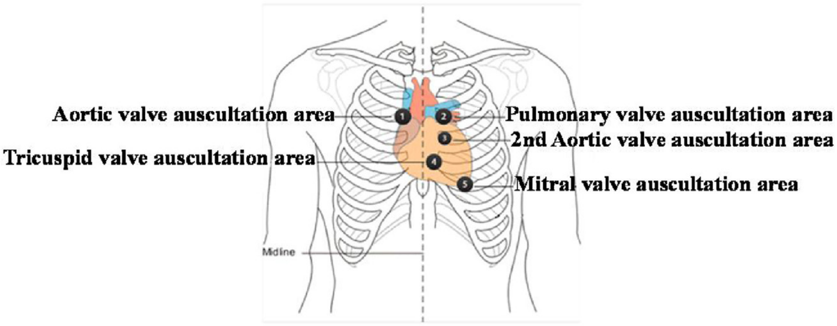
Aortic valve auscultation area
Tricuspid valve auscultation area
Pulmonary valve auscultation area
2nd Aortic valve auscultation area
Mitral valve auscultation area
1
2
3
4
5
Midline

Fig. 6. Locations of heart sound recording.

Table 1
Parameters of the DNN model used in this study.

Parameters	Value
x input (*N*)	78
y output (*M*)	2
Hidden layers (*L*)	3
Neurons in one hidden layer	100
Activation function	Sigmoid

Table 2
Objective assessment matrix.

	Actual class (observation)	
Predicted class (Expectation)	*Tp* (True positive)	*Fp* (False positive) Type I error
	Fn (False negative) Type II error	*Tn* (True negative)

signals may be informed when building the recognition system. With the two parts of noise signals, the noisy training and testing heart sounds were artificially simulated by mixing the clean heart sounds with noise signals at specific SNR levels. The final training set included clean heart sounds and noisy heart sounds at 0, 10, and 20 dB SNR levels. The testing set included six subsets: one subset is the clean heart sounds (from the original test set), and the other five subsets include the noisy heart sounds at 0, 5, 10, 15, and 20 dB SNR levels. Please note that we intentionally designed the training and testing sets to have mismatched SNR conditions, namely 5 dB and 15 dB SNR conditions.

The experimental setup was based on the optimal parameters obtained in our previous study [10]. The sampling frequency was set to 5 kHz (down-sampled from 48 kHz to 5 kHz), 13-dimensional MFCCs were expanded to 39-dimensional ones, and the frame size was set to 15 ms with a 10-ms overlap. The *k*-means algorithm was applied on the MFCC features in a heart sound segment to form a super-vector as the final feature. Table 1 shows the parameters specified in the DNN model.

In this study, we investigated four types of acoustic models, the model trained with original clean data, the model trained with the noisy data, the model trained with the restored data, and the model trained with the combination of clean, noisy, and restored data. Two types of testing data were investigated: the noisy testing data (including the original noise-free testing data) and the restored testing data, which was obtained by applying MMSE-based spectral restoration on the noisy testing data.

4.2. Evaluation metrics

To evaluate the proposed system, we adopted evaluation metrics typically used in pattern recognition and information retrieval, including precision, recall, specificity, and F-measure. To compute these metrics, we consider the four recognition results listed in Table 2. F-measure is also known as F1 measure, which indicates that precision and recall are equal in weight. Recall, which is also called sensitivity, refers to the number of true positive (Tp) instances over the total number of positive instances ($Tp + Fn$). Precision denotes the number of true positive (Tp) instances over all detected instances ($Tp+Fp$). Specificity, also known as true negative rate, refers to the number of true negative (Tn) instances out of the total number of negative instances ($Fp + Tn$). Accuracy denotes the ratio of true positive and true negative instances over the total number of instances evaluated. Equations (16)–(20) show the definitions of these metrics.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{16}$$

$$Recall = \frac{Tp}{Tp + Fn} \tag{17}$$

$$Precision = \frac{Tp}{Tp + Fp} \tag{18}$$

$$Specificity = \frac{Tn}{Fp + Tn} \tag{19}$$

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{20}$$

4.3. Experimental results

This section presents the experimental results, which can be divided into two parts: For the first part, we demonstrate the spectrograms of the heart sound signals as a qualitative analysis. Then, we present the objective analysis results of the MMSE-based signal restoration algorithm using a standardized evaluation metric: speech distortion index (SDI) [18]. The SDI metric has been widely used to perform speech signal analyses, and here we used it to quantitatively analyze the processed heart sound signals.

For the second part, we show the S1 and S2 recognition results. As mentioned earlier, we designed two testing tasks: (1) original heart sounds, denoted as “Test Set-O”; and (2) MMSE restored heart sounds, denoted as “Test Set-R”. For each task, we tested recognition using four acoustic models, which were trained on (1) clean training data, denoted as the “clean model”; (2) noisy training data, denoted as the “MT model”; (3) restored training data, denoted as the “SR model”; (4) spectral restoration with multi-style training data, denoted as the “SRMT model”. It is clear that the training set for training SRMT model was larger than the other three sets. In order to have a fair comparison, we duplicated the training samples for the other three sets, and thus the four training sets contained the same amount of training data. For all of four models, the DNN model had the same structure as reported in Table 1. Accuracy was used as the main evaluation metric to test these models.

Table 3
SDI values of the MMSE-based spectral restoration algorithm (denoted as "MMSE") on the testing data. SDI values of the non-processed original data (denoted as "Original") are also listed for comparison.

	20 dB	15 dB	10 dB	5 dB	0 dB
Original	14.77	15.07	16.29	20.61	35.13
MMSE	0.44	0.49	0.58	0.88	1.10

Table 4
Recognition accuracies of clean, MT, SR, and SRMT acoustic models tested on the Test Set-O.

Condition	Clean	MT	SR	SRMT
Clean	87.40	88.18	87.87	88.22
20 dB	86.30	88.07	88.20	88.61
15 dB	86.08	87.47	87.73	87.69
10 dB	85.33	87.20	87.73	87.76
5 dB	84.71	86.54	87.69	87.84
0 dB	82.83	85.72	86.50	86.70
Average	85.44	87.20	87.62	87.80

4.3.1. Qualitative heart sound signal analysis

Fig. 7(a) and (b) show the spectrograms of clean heart sound and pure noise, and Fig. 7(c) and (d) show the noisy sound at 0 dB SNR level and the restored heart sound, respectively. These spectrograms can display how the frequencies present in a heart sound signal varying over time [46].

From Fig. 7(a)–(d), we can first note that the noise components may considerably blur the acoustic patterns of the clean heart sound, as shown in Fig. 7(c). Furthermore, we note that the noise components can be effectively removed by the MMSE spectral restoration algorithm, as shown in Fig. 7(d), where the heart sound signals can be easily noted.

4.3.2. Quantitative heart sound signal analysis

The SDI metric corresponds to the ratio of the energies of the residual signal and clean speech signals:

$$\mathrm{SDI} = \frac{E\left[\left(s[n]-\hat{s}[n]\right)^2\right]}{E\left[s^2[n]\right]} \tag{21}$$

where $\left(s[n]-\hat{s}[n]\right)$ denotes the residual signal, and $s[n]$ denotes the clean signal. A lower SDI indicates a smaller difference between the clean signal and the target signal to be compared with. In Table 3, we list the SDI results of the MMSE spectral restoration method (where $s[n]$ and $\hat{s}[n]$ are clean signal and the restored signal, respectively). The SDI results of the original noisy signals are also presented for comparison (where $s[n]$ and $\hat{s}[n]$ are clean signal and the noisy signal, respectively). The testing data were used to compute the SDI values in Table 3.

From the results in Table 3, we can easily note the differences of noisy and clean heart sounds: the SDI results for 20 dB to 0 dB monotonously increased from 14.77 to 35.13. On the other hand, the MMSE algorithm considerably reduced the differences: the SDI results for 20 dB to 0 dB ranged from 0.44 to 1.10. The results confirm the effectiveness of the MMSE to reduce the noisy components and to restore the clean signals.

4.3.3. Recognition results on Test Set-O (original heart sounds)

In this section, we tested recognition on Test Set-O, including clean, 0, 5, 10, 15, and 20 dB SNR levels of heart sounds. In our preliminary experiments, we noted that the DNN training model with randomly starting values may lead to different results for each test. Therefore, we ran each experiment 25 times and reported the average of 25 results for each testing condition. Table 4 lists the testing results of clean, MT, SR, and SRMT models.

Table 5
Recognition accuracies of clean, MT, SR, and SRMT acoustic models tested on the Test Set-R.

Condition	Clean	MT	SR	SRMT
Clean	78.87	76.84	79.09	82.14
20 dB	81.52	82.25	87.12	88.62
15 dB	80.82	81.02	86.96	86.87
10 dB	80.51	79.29	84.75	84.97
5 dB	78.83	76.07	85.08	86.5
0 dB	70.81	69.44	76.27	78.74
Average	78.56	77.49	83.21	84.64

From Table 4, we first note that when the clean acoustic model was used, the accuracy of the noise-free testing condition reached 87.40%. However, as the noise increased, the accuracy decreased accordingly, culminating with a mean of 82.83% at the 0 dB SNR condition. These results show that the noise involvements notably reduced the recognition accuracy. Next, when compared with the clean model, the MT model can give higher accuracies, for both clean and noisy testing conditions, suggesting that noisy training can increase the stability of the model when dealing with noisy testing data.

Moreover, the results in Table 4 also show that the SR model yielded higher accuracies than the clean model while similar performance as compared to the MT model, showing that both MT and SR can be used to improve model robustness against noise interferences. Finally, the SRMT model yielded a similar performance to the MT and SR models under noise-free testing conditions, and demonstrated better average recognition performance among the four models in the range of 0–20 dB SNR noisy conditions. The results confirm that the proposed SRMT is the best option under both clean and noisy conditions, given the original test data

4.3.4. Recognition results on Test Set-R (restored heart sounds)

Next, we compare the testing results of the clean, SR, MT, and SRMT acoustic models tested on Test Set-R. The results are presented in Table 5. When comparing the results of Tables 4 and 5, we note that the recognition accuracy dropped significantly across different testing conditions. The results show that the online spectral restoration may not effectively enhance recognition accuracies. On the other hand, the offline SRMT-based training can build an acoustic model that is robust to noise interferences, as shown in Table 4.

In addition to recognition accuracies, we used a statistical hypothesis test, the dependent *t*-test [47,48], to verify the significance of performance improvements. The *t*-test was used to compare two methods by testing the matched-pair accuracy results, and the p-values were used to report the *t*-test results. Small p-values suggested consistent improvements of method-II over method-I. We compared the results of clean and SRMT in Table 4 and Table 5, and the corresponding p-values are 0.0015 and 0.00025, respectively. With such small p-value, we confirm that the results of the SRMT model significantly outperformed the one of the clean model.

4.3.5. Recognition results of other evaluation metrics

From the results of Tables 4 and 5, we note that the SRMT model tested on the original test data demonstrates better performance. The best results are achieved using the MMSE-based spectral restoration technique to synthesize more training data, with no usage of additional online spectral restoration process. In this section, we provide other evaluation metrics of clean and SRMT models to further demonstrate the effectiveness of the SRMT approach. Fig. 8 presents the average precision, recall, specificity, and F-measure of S1 over 0–20 dB SNR conditions for the Clean and SRMT models, both tested with original test data. Because the S1

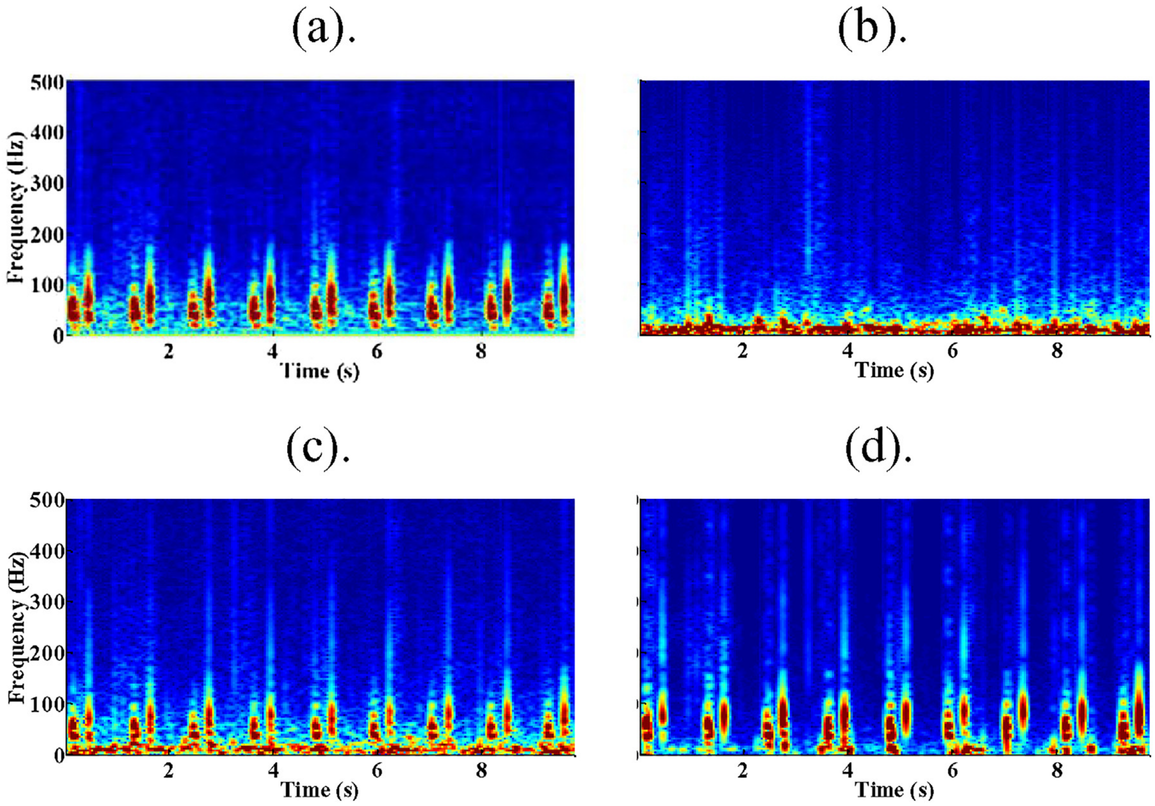

Fig. 7. Spectrograms in 0–500 Hz of (a) clean heart sound, (b) pure noise, (c) noisy heart sound at 0 dB SNR, and (d) restored heart sound with the MMSE-based spectral restoration.

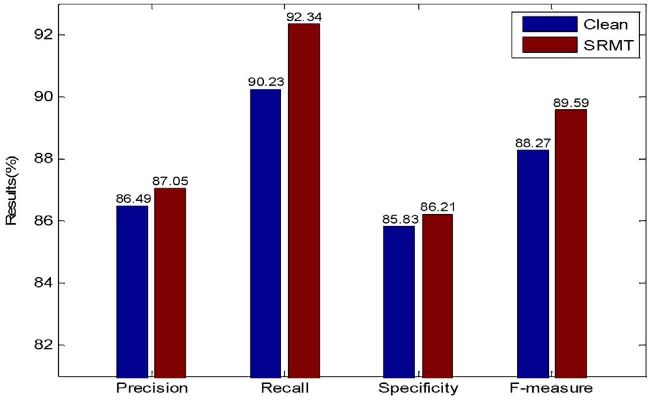

Fig. 8. Precision, recall, specificity, and F-measure scores of S1 obtained using Clean and SRMT on original test data.

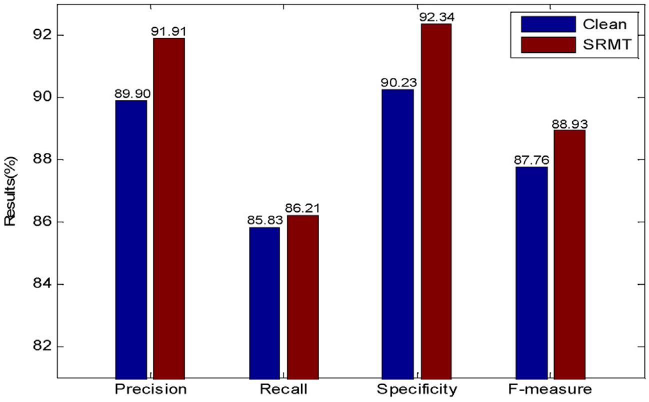

Fig. 9. Precision, recall, specificity, and F-measure scores of S2 obtained using Clean and SRMT on original test data.

and S2 recognition is a binary classification task, the S1 specificity equals the S2 recall, and the S2 specificity equals the S1 recall. From the results in Figs. 8 and 9, we note that SRMT model achieves higher average precision, recall, specificity, and F-measure than the clean model.

We also list the average precision, recall, specificity, and F-measure of S2 over 0–20 dB SNR conditions for the clean and SRMT models, both tested with original test data in Fig. 9. Similar findings as those observed in Fig. 8 are also noted, again showing the effectiveness of the combination of SR and MT for improving performance robustness in noise.

5. Conclusion

This study proposed a spectral restoration with multi-style training (SRMT) approach to improve the model robustness against noise interferences. The experimental results confirmed that the presented SRMT method yields higher accuracy than the clean model, confirming that through the SRMT approach, the acoustic models can be trained to be robust under noisy conditions. We also used two different test sets, original testing data and MMSE-based restored testing data. The results show that the models trained by SRMT yield higher accuracy with the original test data than that with the MMSE-based restored testing data. The findings from this study show that by incorporating the restored data, the model can be trained to be robust against noise interferences, and no spectral restoration operations are required online. Moreover, the proposed SRMT method can compensate for mismatch issues and maintain system stability. In the future, we will test the proposed SRMT using more challenging noise types. Meanwhile, we will try to incorporate deep learning based spectral restoration techniques into the SRMT framework.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[4] H. Liang, S. Lukkarinen, I. Hartimo, Heart sound segmentation algorithm based on heart sound envelogram, Proceedings Computers in Cardiology (1997) 7–10.

[5] D. Kumar, P. Carvalho, M. Antunes, J. Henriques, L. Eugenio, R. Schmidt, J. Habetha, Detection of S1 and S2 heart sounds by high frequency signatures, Proceedings EMBS (2006) 1410–1416.

[6] J.E. Hebden, J.N. Torry, Neural network and conventional classifiers to distinguish between first and second heart sounds, Proceedings Artificial Intelligence Methods for Biomedical Data Processing, IEE Coll. vol. 3 (1996) 1–6.

[7] A.C. Stasis, E.N. Loukis, S.A. Pavlopoulos, D. Koutsouris, Using decision tree algorithm as a basis for a heart sound diagnosis decision support system, Proceedings Artificial Intelligence Methods for Biomedical Data Processing, IEE Coll. (1996) 1–6.

[8] T. Olmez, Z. Dukar, Classification of heart sounds using an artificial neural network, Pattern Recognit. Lett. 24 (2003) 617–629.

[9] D. Kumar, P. Carvalho, P. Gil, J. Henriques, M. Antunes, L. Eugenio, A new algorithm for detection of S1 and S2 heart sounds, in: Proceedings ICASSP, 2006, pp. 1180–1183.

[10] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, C.-C. Wu, S1 and S2 heart sound recognition using deep neural networks, IEEE Trans. Biomed. Eng. 64 (2017) 372–380.

[11] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[12] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Sparse representation based on a bag of spectral exemplars for acoustic event detection, Proc. ICASSP (2014) 6255–6259.

[13] H. Liang, S. Lukkarinen, I. Hartimo, Heart sound seg-mentation algorithm based on heart sound envelogram, Proc. CinC (1997) 105–108.

[14] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl, Constrained K-means clustering with background knowledge, Proceedings Machine Learning (2001) 577–584.

[15] J.S. Lim, A.V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proceedings of the IEEE 67 (12) (1979) 1586–1604.

[16] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. 27 (2) (1979) 113–120.

[17] J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, Adaptive β-order generalized spectral subtraction for speech enhancement, Signal Processing 88 (11) (2008) 2764–2776.

[18] J. Chen, J. Benesty, Y.A. Huang, E.J. Diethorn, Fundamentals of noise reduction, in: Springer Handbook of Speech Processing, Springer, 2008, pp. 843–872.

[19] R. McAulay, T. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, IEEE Trans. Acoust. 34 (no. 4) (1986) 744–754.

[20] R.J. McAulay, Shape invariant time-scale and pitch modification of speech, in: T.F. Quatieri (Ed.), IEEE Trans. Signal Process., 40, 1992, pp. 497–510.

[21] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (3) (1975) 561–580.

[22] S. Suhadi, C. Last, T. Fingscheidt, A data driven approach to a priori SNR estimation, IEEE Trans. Audio Speech Lang. Process. 19 (1) (2011) 186–195.

[23] T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model, EURASIP J. Appl. Signal Processing 2005 (2005) 1110–1126.

[24] U. Kjems, J. Jensen, Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement, Proc. EUSIPCO (2012) 295–299.

[25] R. McAulay, M. Malpass, Speech enhancement using a soft decision noise suppression filter, IEEE Trans. Acoust. 28 (2) (1980) 137–145.

[26] Y. Tsao, Y.-H. Lai, Generalized maximum a posteriori spectral amplitude estimation for speech enhancement, Speech Commun. 76 (2016) 112–126.

[27] B. Li, Y. Tsao, K.C. Sim, An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition, Proc. INTERSPEECH (2013) 3002–3006.

[28] T. Ko, V. Peddinti, D. Povey, M. Seltzer, S. Khudanpur, A study on data augmentation of reverberant speech for robust speech recognition, Proc. ICASSP (2017).

[29] X. Cui, V. Goel, B. Kingsbury, Data augmentation for deep neural network acoustic models, IEEEACM Trans. Audio Speech Lang. Process. 23 (9) (2015) 1469–1477.

[30] P. Lin, D.-C. Lyu, F. Chen, S.-S. Wang, Y. Tsao, Multi-style learning with denoising autoencoders for acoustic modeling in the internet of things (IoT), Comput. Speech Lang. 46 (2017) 481–495.

[31] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T.F. Zheng, Y. Li, “Noisy training for deep neural networks in speech recognition, Eurasip J. Audio Speech Music. Process. (2015) 1–14.

[32] I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. 11 (5) (2003) 466–475.

[33] R.P. Lippmann, E.A. Martin, D.B. Paul, Multi-style training for robust isolated-word speech recognition, Proc. ICASSP (1987) 705–708.

[34] M.L. Seltzer, D. Yu, Y. Wang, An investigation of deep neural networks for noise robust speech recognition, Proc. ICASSP (2013) 7398–7402.

[35] C. Weng, D. Yu, M.L. Seltzer, J. Droppo, Deep neural networks for single channel multi-talker speech recognition, IEEEACM Trans. Audio Speech Lang. Process. 23 (10) (2015) 1670–1679.

[36] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, Robust Automatic Speech Recognition: a Bridge to Practical Applications, Academic Press, 2015.

[37] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoustic Speech Signal Process. ASSP-29 (1981) 254–272.

[38] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE Trans. Audio Speech Lang. Process. 23 (2015) 7–19.

[39] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Speech enhancement based on deep denoising autoencoder, Proc. Inter-Speech (2013).

[40] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (2012) 82–97.

[41] S.M. Siniscalchi, T. Sveendsen, C.-H. Lee, An artificial neural network approach to automatic speech processing, Neurocomputing (2014) 326–338.

[42] S.M. Siniscalchi, J. Li, C.-H. Lee, Hermitian polynomial for speaker adaptation of connectionist speech recognition systems, IEEE Trans. Audio Speech Lang. Process. 21 (10) (2013) 2152–2161.

[43] P.Y. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: Proc. ICDAR, 2003, pp. 958–963.

[44] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: Proc. CVPR, 2012, pp. 3642–3649.

[45] Please refer to FDA website https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K160023.

[46] J.L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972.

[47] A.J. Hayter, Probability and Statistics for Engineers and Scientists, 3rd ed., Duxbury Press, 2006.

[48] A. Agresti, C.A. Franklin, Statistics: the Art and Science of Learning from Data (MyStatLab Series), Prentice Hall, 2008.