This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

# S1 and S2 Heart Sound Recognition using Deep Neural Networks

*Tien-En Chen, Shih-I Yang, Li-Ting Ho, Kun-Hsi Tsai, Yu-Hsuan Chen, Yun-Fan Chang, Ying-Hui Lai, Syu-Siang Wang, Yu Tsao*, and Chau-Chung Wu*

***Abstract*—*Objective*: This study focuses on the first (S1) and second (S2) heart sound recognition based only on acoustic characteristics; the assumptions of the individual durations of S1 and S2 and time intervals of S1–S2 and S2–S1 are not involved in the recognition process. The main objective is to investigate whether reliable S1 and S2 recognition performance can still be attained under situations where the duration and interval information might not be accessible. *Methods*: A deep neural network (DNN) method is proposed for recognizing S1 and S2 heart sounds. In the proposed method, heart sound signals are first converted into a sequence of Mel-frequency cepstral coefficients (MFCCs). The K-means algorithm is applied to cluster MFCC features into two groups to refine their representation and discriminative capability. The refined features are then fed to a DNN classifier to perform S1 and S2 recognition. We conducted experiments using actual heart sound signals recorded using an electronic stethoscope. Precision, recall, F-measure, and accuracy are used as the evaluation metrics. *Results*: The proposed DNN-based method can achieve high precision, recall, and F-measure scores with more than 91% accuracy rate. *Conclusion*: The DNN classifier provides higher evaluation scores compared with other well-known pattern classification methods. *Significance*: The proposed DNN-based method can achieve reliable S1 and S2 recognition performance based on acoustic characteristics without using an ECG reference or incorporating the assumptions of the individual durations of S1 and S2 and time intervals of S1–S2 and S2–S1.**

***Index Terms*—Heart sound recognition, deep neural networks, acoustic fingerprinting, S1 and S2 recognition**



Tien-En Chen is with the Division of Cardiology, Department of Internal Medicine, China Medical University Hospital, Taichung, Taiwan; Medical College, China Medical University, Taichung, Taiwan (e-mail: chentienen@mail.cmuh.org.tw).

Shih-I Yang is with Department of Emergency Medicine, Everan Hospital, Taichung, Taiwan (e-mail: 00032@everanhospital.com.tw).

Li-Ting Ho and Chau-Chung Wu are with National Taiwan University College of Medicine, Taipei, Taiwan (e-mails: {100489, chauchungwu}@ntu.edu.tw).

Kun-Hsi Tsai is with iMediPlus Inc., Hsinchu, Taiwan (e-mails: Peter.tsai@imediplus.com).

Yu-Hsuan Chen is with Division of Chest Medicine, Department of Internal Medicine, Cheng Hsin General Hospital, Taipei, Taiwan (e-mail: ch1571@chgh.org.tw).

Yun-Fan Chang, Ying-Hui Lai, Syu-Siang Wang, and Yu Tsao* are with the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei, Taiwan (e-mails: {she213, jackylai, sypdbhee, yu.tsao}@citi.sinica.edu.tw ).

## I. Introduction

Cardiac auscultation is a conventional physical examination for evaluating cardiac function and activities. Audible heart sounds are generated by cardiac valves snapping shut or by turbulent flows. In healthy adults, two normal heart sounds occur in sequence in a cardiac cycle. The pitch and occurrence time of heart sounds follow certain patterns. The first heart sound (S1) occurs at the start of the ventricular systole phase, which results from closing the mitral and tricuspid valves (collectively known as the atrioventricular valves). The second heart sound (S2) occurs at the start of the ventricular diastole phase, which results from closing the aortic and pulmonic valves. S1 is a low-pitch sound with longer duration, whereas S2 is a high-pitch sound with shorter duration. In normal situations, the S1–S2 interval (systole) is shorter than the S2–S1 interval (diastole).

Numerous heart diseases can be diagnosed effectively through auscultation [1]. In some fatal heart diseases, such as acute valvular dysfunction, cardiac auscultation has been confirmed to be successful, reliable, and inexpensive for early diagnosis. However, auscultation with a stethoscope has low diagnostic sensitivity and accuracy in imperfect acoustic environments, and the accuracy of cardiac auscultation is doctor and experience-dependent [2]. Moreover, some diseases demonstrate special occurrence patterns. Thus, how to automatically detect and accurately evaluate the occurrence time of S1 and S2 has become a crucial task. This task can effectively help doctors diagnose the occurrence of diseases [3].

Phonography (PCG) is a method for recording heart sounds by collecting mechanical vibrations generated from heart beats or turbulent flows with a stethoscope on various parts over the chest wall; in this process, recorded heart sounds are presented graphically [4]. PCG could provide quantitative and qualitative information of heart sounds and murmurs. Studies on heart sound detection can be divided into two categories: ECG signal-dependent and ECG signal-independent. ECG signal-dependent studies include ECG-based instantaneous energy detections [5] and R wave and T wave detection [6], [7]. In actual clinical cases, however, simultaneously and synchronously recording and analyzing ECG and PCG measurements is not always practical. Thus, ECG signal-independent (using only PCG) has become an important research topic. Previous ECG signal-independent studies used signal processing (e.g., normalized average Shannon energy [8], high-frequency-based methods [9]), and statistical modeling approaches (e.g., neural network (NN) classifiers [10] and decision trees [11]). In addition, some approaches have proposed incorporating the assumptions of S1–S2 and S2–S1 intervals to further improve classification performance [12], [13]. Although using interval

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

regularity can improve performance in normal situations, these approaches might not be applicable for patients with arrhythmia.

This paper proposes a novel acoustic fingerprinting-based detection framework that applies only supervised classifiers for recognizing S1 and S2. Mel-frequency cepstral coefficients (MFCC) [14], typically used for classifying acoustic incidents, are applied for feature extraction. Subsequently, the K-means algorithm [15] is used to divide one heart sound fragment into two groups. A center feature vector is formed for each of the two groups, and these two feature vectors are concatenated to form a super-vector. Finally, the super-vector is fed into a deep NN (DNN) classifier [16]–[20] in order to classify S1 and S2. To demonstrate the effectiveness of DNN for S1 and S2 recognition, we compare it with other well-known classifiers, namely K-nearest neighbor (KNN) [21], Gaussian mixture model (GMM) [22], logistic regression (LR) [23], and support vector machine (SVM) [24]. The experiment results indicate that the proposed DNN classifier exhibits higher precision, recall, F-measure, and accuracy in recognizing S1 and S2 compared with existing methods.

The rest of this paper is organized as follows. Section II presents related works and the overall architecture for the proposed S1 and S2 recognition system. Sections III and IV correspondingly introduce the feature extraction and DNN classifier used in the proposed system. Section V presents the experiment procedures, including an introduction to the experiment data set, experiment setup, and evaluations of precision, recall, F-measure, and accuracy. Finally, section VI presents the conclusion of this study.

## II. Related Works and Proposed S1 and S2 Recognition System

This section first presents related works on heart sound recognition and discusses the major advantages of the proposed system. Then, we introduce the overall system architecture.

### A. Related works

Numerous signal processing and machine learning algorithms have been developed to perform S1 and S2 heart sound recognition. In [9], a high frequency marker was developed by applying the fast wavelet transform (FWT) on heart sound signals to extract the high frequency signatures of the valve closing sounds. By properly incorporating the information on S1–S2 and S2–S1 duration, the proposed system could achieve excellent performance in the case of arrhythmic heart sounds [9]. Meanwhile, an S1 and S2 heart sound identification system was proposed that uses the Mel-filter coefficient features and self-organizing-map (SOM) model. With a post-processing stage designed based on heart sound duration and interval, the recognition results were comparable to those obtained using ECG signals [13]. In [10], time intervals and diastolic periods of heart sounds were used directly to devise features that were then inputted into an artificial NN (ANN) and several other classifiers. The experiment results showed that ANN provides higher classification rates when compared with other approaches [10]. Olmez and Dukar proposed a three-stage heart sound classification system [12]. In the system, the heart sound segments were first detected by finding the peaks of continuous segments that exceed the threshold limit. Then, the time interval and diastolic period information were used to perform S1 and S2 identification.

The present study has three major goals: (1) Unlike previous studies, we intend to explore the performance of S1 and S2 recognition based only on acoustic fingerprinting without incorporating additional duration and interval information. (2) Based on the recent success of deep learning in acoustic modeling [18], [20], we build an acoustic model of S1 and S2 heart sounds using DNN for the first time (to the best of our knowledge). (3) We investigate the effectiveness of several advanced deep learning algorithms on this heart sound recognition task.

### B. Overall S1 and S2 recognition architecture

Fig. 1 shows the flowchart for the proposed S1 and S2 recognition framework. As shown in this figure, the feature extraction process is first performed to convert heart sound signals into a set of feature vectors. In the offline phase, the feature vectors for S1 and S2 are used to train a DNN classifier. In the online phase, testing feature vectors are introduced into the DNN classifier, and the testing data category is determined according to the classifier output.

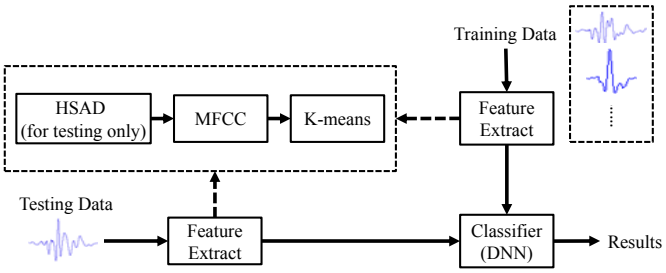


Fig. 1. Flowchart for proposed S1 and S2 recognition system.

## III. Feature Extraction for Proposed System

This section details the feature extraction process shown in Fig. 1. As shown in this figure, the feature extraction comprises heart sound activity detection (HSAD), MFCC, and K-means processes. In this section, we present the procedures for MFCC and K-means. The HSAD process is presented in Section V.

### A. MFCC

The MFCC feature extraction procedure comprises six operations: pre-emphasis, windowing, fast Fourier transform (FFT), Mel-filtering, nonlinear transformation, and discrete cosine transform (DCT). The pre-emphasis operation enhances the received signals to compensate for signal distortions. The windowing operation divides a given signal into a sequence of frames. The FFT operation is applied to the windowed signals for spectral analysis. The Mel-filtering operation is designed based on human perception, and it integrates the frequency compositions from one Mel-filter band into one energy intensity. The non-linear transformation operation takes the logarithm of all Mel-filter band intensities. The transformed intensities are then converted into MFCCs using DCT. MFCC feature extraction has been confirmed to be effective in speech recognition [25], speaker recognition [26], and various acoustic pattern recognition tasks [27], [28].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

For numerous applications, using differential parameters to describe temporal characteristics improves pattern recognition performance. A differential cepstral parameter is defined as the slope of a cepstral parameter versus time, representing the dynamic change of the cepstral parameters. Hence, three times of dimensions of original features can be obtained after appending velocity and acceleration features [26]. Velocity ($vel$) and acceleration ($acc$) features can be calculated as follows:

$$vel(d,t) = \frac{\sum_{m=1}^{M_v} m \times [c(d,t+m) - c(d,t-m)]}{2\sum_{m=1}^{M_v} m^2} \quad (1)$$

$$acc(d,t) = \frac{\sum_{m=1}^{M_a} m \times [vel(d,t+m) - vel(d,t-m)]}{2\sum_{m=1}^{M_a} m^2} \quad (2)$$

where $c(d,t)$ is the $d$th dimension of the cepstral parameter, and $t$ is the time indicator for the current sound frame; $M_v$ and $M_a$ are window lengths for computing *vel* and *acc* coefficients, respectively. In this study, we set $M_v = 3$ and $M_a = 2$.

### *B. K-means algorithm*

The main goal of the K-means algorithm is to determine representative data points from large numbers of data points. Such data points are called "population centers." Data compression (i.e., using a low number of data points to represent a high amount of data for compressing data) and classification (i.e., using a low number of representative points to represent specific categories for lowering the amount of data and avoiding adverse effects caused by noise) are applied to population centers. The calculation steps of the K-means algorithm are presented as follows:

a. Initialization:
   Divide training materials $v_i, i = 1, \dots, N$, randomly into $K$ groups and arbitrarily choose one observation from each group as the initial population center $\mu_k$, $k$ = 1,2,…,$K$.
b. Recursive calculation:
   i. Let each $v_i$ find the nearest population center and assign it to that population center by
      $$k^* = \arg_k \min d(v_i, \mu_k), i = 1, \dots, N \quad (3)$$
      where $d(\cdot,\cdot)$ denotes the distance measure.
   ii. All $v_i$ that belong to the $k$-th group form a group. Calculate the population center $\mu_k$ again.
   iii. If the new groups of the population centers are the same as the original population center set, training is completed. Otherwise, new population groups replace the original population center groups. Step a. is repeated to continue recursive calculations.

In this study, the K-means algorithm is used to cluster the acoustic features within each heart sound segment into two groups ($K$ = 2). Then a population center vector is computed for each group. These two center vectors are then concatenated to form a super-vector. This super-vector is the final feature that represents a segment of heart sound. The super-vectors are used to build classifiers and perform S1/S2 recognition.

## IV. DNN Classifier for Proposed System

ANN is a mathematical model that mimics biological NN structures and allows a computer system to execute classification or regression tasks. Numerous scholars have proposed diverse NN models for solving various problems. More recently, NNs with multiple layers (known as DNN) have shown outstanding performance in a wide variety of tasks, such as automatic speech recognition [18], speech processing [29], [30], and visual pattern recognition [31], [32]. The operating principle of DNN [17], [18] involves using the current output layer as the input for the next hidden layer. The concept of this algorithm is to use a high number of hidden layers to strengthen the classification or regression capability.

Fig. 2 shows the structure of a DNN model. The relationship between the input feature, $\boldsymbol{x}$, and output of the first hidden layer, $\boldsymbol{a}_1$, is described as follows:

$$\boldsymbol{a}_1 = F(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \quad (4)$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ respectively correspond to the weight matrix and bias vector in the first layer, and $F(.)$ is the activation function. After obtaining the output of the first hidden layer, the relationship between the current and the next hidden layer can be expressed as follows:

$$\boldsymbol{a}_l = F(\boldsymbol{W}_l\boldsymbol{a}_{l-1} + \boldsymbol{b}_l), l = 2, \dots, L, \quad (5)$$

where $L$ is the total number of layers in the DNN model. Another function, $G(.)$, is usually applied on the output layer, which is used to execute classifications or regressions. Thus, we have

$$\hat{\boldsymbol{y}} = G(\boldsymbol{a}_L) \quad (6)$$

where $\hat{\boldsymbol{y}}$ is the DNN output. For classification tasks, the softmax function is generally used for $G(.)$ [33]. Given the correct label, $\boldsymbol{y}$, the parameters of the DNN classifier can be estimated as follows:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \{C(\boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{x}, \boldsymbol{\theta}) + \gamma R(\boldsymbol{\mathcal{W}}) + \eta\rho(\boldsymbol{\mathcal{A}})\} \quad (7)$$

where $\boldsymbol{\theta} = \{\boldsymbol{W}_l, \boldsymbol{b}_l, l = 1,2, \dots, L\}$ is the DNN parameter set and $C(.)$ is a cost function.

In this study, cross-entropy [34] is used as the cost function. In order to train the DNN model, we prepare a set of training data, $\boldsymbol{X} = [\boldsymbol{x}_1, \dots \boldsymbol{x}_i, .. \boldsymbol{x}_N]$, and the corresponding output labels $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots \boldsymbol{y}_i, .. \boldsymbol{y}_N]$, where $N$ is the total number of training samples; thus, the cost function becomes

$$C(\boldsymbol{Y}, \hat{\boldsymbol{Y}}; \boldsymbol{X}, \boldsymbol{\theta}) = \frac{-1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}[y_{i,j} log \hat{y}_{i,j}], \quad (8)$$

where $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_1, \dots \hat{\boldsymbol{y}}_i, .. \hat{\boldsymbol{y}}_N]$ is the DNN output ($\hat{\boldsymbol{y}}_i$ is the $i$th DNN output given input $\boldsymbol{x}_i$); $y_{i,j}$ and $\hat{y}_{i,j}$ denote the $j$th element of $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$, respectively.

In Eq. (7), $R(\boldsymbol{\mathcal{W}})$ is estimated as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

$$R(\boldsymbol{W}) = \sum_{l} \|\boldsymbol{W}_l\|_F^2 \tag{9}$$

where $\|.\|_F^2$ denotes the Frobenius norm, and $\rho(\boldsymbol{\mathcal{A}})$ is the sparsity penalty of the hidden outputs [35], [36]; $\gamma$ and $\eta$ are the controlling coefficients. These two regularizations are used at the DNN training stage to mitigate over-fitting. According to Eqs. (7)-(9), the standard back-propagation algorithm is applied to compute the parameters in the DNN model [19], [20].

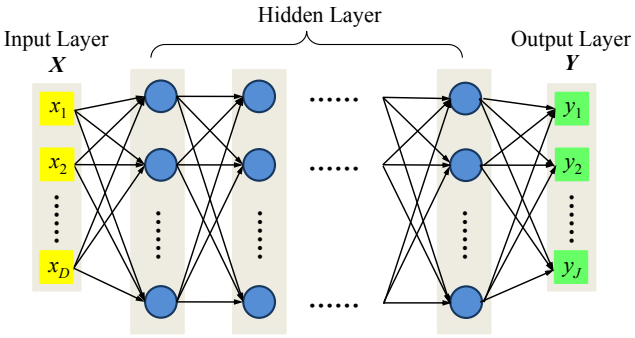


Fig. 2. Structure of a DNN model.

The advantage of DNN is that it has a deep and complex structure. An intuitive limitation is that when the training data is insufficient, the DNN parameters ($\boldsymbol{\theta}$) might not be estimated accurately. To overcome this limitation, a pre-training technique that uses unlabeled data is generally adopted [20], [37]. A deep belief network (DBN) and maximum likelihood (ML) criterion are involved in the pre-training process. A DBN model is formed by stacking a set of restricted Boltzmann machine (RBM) models. Given the training data, $\boldsymbol{z}$, the RBM model defines a joint probability function as follows:

$$p(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda}) = \frac{\exp(-E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda}))}{\sum_v \sum_h \exp(-E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda}))} \tag{10}$$

where $\boldsymbol{h}$ and $\boldsymbol{\Lambda}$ denote the hidden units and parameter set (including wright and bias terms), and $E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda})$ is an energy function. For Bernoulli (visible)-Bernoulli (hidden), the energy function is given as follows:

$$\mathrm{E}(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda}) = -\sum_{i=1}^{Z}\sum_{j=1}^{H} w_{ij} z_i h_j - \sum_{i=1}^{Z} b_i z_i - \sum_{j=1}^{H} a_j h_j \tag{11}$$

where $Z$ and $H$ are the numbers of visible and hidden units. Similarly, for Gaussian (visible)-Bernoulli (hidden), the energy function is given as follows:

$$\mathrm{E}(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\Lambda}) = -\sum_{i=1}^{Z}\sum_{j=1}^{H} w_{ij} z_i h_j + \frac{1}{2}\sum_{i=1}^{Z} (z_i - b_i)^2 - \sum_{j=1}^{H} a_j h_j \tag{12}$$

The contrastive divergence (CD) algorithm is generally adopted for estimating the parameters in RBMs [20].

When estimating a DBN model, the parameters of the first RBM are estimated using the training data, $\boldsymbol{z}$. In the following layers, the greedy learning algorithm is adopted to train another RBM using the hidden activations of the previous layer as the input data. This process continues until it reaches the last layer and forms a DBN. Then, a softmax function is added to the top of the DBN model, and the standard back-propagation training with the cost function in Eq. (7) is applied to estimate the DNN parameters.

## V. Experiments

In the experiment, the feature extraction process as presented in Section III was collocated with the DNN model as presented in Section IV in order to build a classification system. In the training stage, S1 and S2 training data were manually segmented and labeled by a medical doctor. The S1 and S2 segments were then collected and used to train a DNN classifier. In the testing phase, the heart sounds were classified by the trained DNN model. In this section, we first present the experiment setup and evaluation metrics. Next, the experiment results are presented. In the experiments, we first investigate the optimal setups for acoustic features and DNN structures. In order to evaluate the performance of the DNN classifier, KNN, LR, SVM, and GMM classifiers were also implemented and recognition was tested for comparison.

### *A. Experiment setup*

Actual audio data collected by our research team were used in this experiment. The procedures were reviewed and approved by the local institutional review board (IRB) committees. The objective of this study is to recognize heart sounds based on their acoustic characteristics. Fig. 3 illustrates a signal processing block diagram of the electronic stethoscope used in this study.

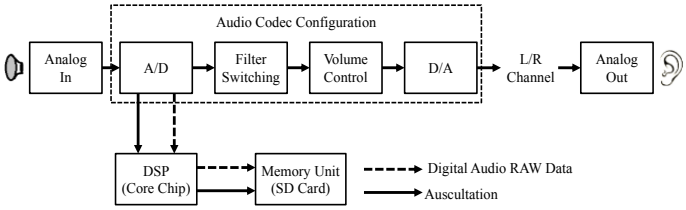


Fig. 3. Block diagram for electronic stethoscope used in study.

As shown in Fig. 3, the Analog In unit receives the incoming sounds, the A/D Sampling unit converts the signal from analog to digital form at a 48-kHz sampling rate, the Filter Switching unit emphasizes sounds in the lower frequency bands, the Volume Control unit adjusts the gain of the output sounds, the D/A unit reconstructs the signal from digital to analog, and the Analog Output unit outputs the reconstructed sounds. The DSP unit processes the digital signal for storage in the Memory unit, and the Memory unit stores raw data that can be used for further analysis in the future.

We recorded three sets of audio data to form training, development, and testing sets. Audio recordings of 16 people (11 males and 5 females) constituted the training set, which includes a total of 313 S1 heart sounds and 313 S2 heart sounds. The development set comprised the recordings of six people (three males and three females), and 60 S1 sounds and 60 S2 sounds were obtained. Audio recordings of six people (three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

males and three females) constituted the testing set, which includes 87 S1 heart sounds and 87 S2 heart sounds. Notably, the data for the training, testing, and development sets were mutually exclusive. For the recordings that belong to the training and development sets, S1 and S2 sounds were manually segmented; on the other hand, for those that belong to the testing set, an HSAD process was performed to detect heart sound segments before performing recognition. Fig. 4 illustrates the heart sound recording positions, which were centralized on the second left intercostal space along the left sternal border ① and third intercostal space on the left sternal border ②.

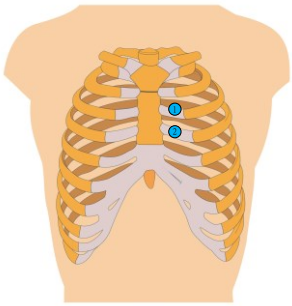


Fig. 4. Positions of heart sound recording in this study.

## B. Evaluation metrics

The evaluation metrics typically used in pattern recognition and information retrieval are precision, recall, and F-measure. These metrics serve as the standards for evaluating the effectiveness of a pattern recognition system [38]. Table I lists the four items used in the evaluations.

TABLE I
EVALUATION METRICS

| | Actual Class (Observation) | |
|---|---|---|
| Predicted Class (Expectation) | Tp (True positive) Correct result | Fp (False positive) False alarm |
| | Fn (False negative) Missed detection | Tn (True negative) Correct absence of result |

Equations (13)–(15) show the definitions of these metrics.

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{13}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \tag{14}$$

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

F-measure is also called $F_1$ measure, and it represents the equal weights of precision and recall. Recall is typically denoted as true positive rate or sensitivity, and precision is denoted as a positive predictive value. Accuracy is generally used as the evaluating metric for classification processes:

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{16}$$

## C. Experiment results

Fig. 5 depicts the S1 and S2 spectrogram and waveform. A spectrogram is often used to display how the frequencies present in a temporal signal that varies over time [39]. To draw the spectrogram shown in Fig. 5, the sampling frequency was set to 5 kHz, and the frame size was set to 20 ms, with a 10-ms overlap. From the spectrogram, it can be noted that the S1 and S2 heart sounds are mainly concentrated in the low-frequency regions. Moreover, S1 has longer duration than S2.

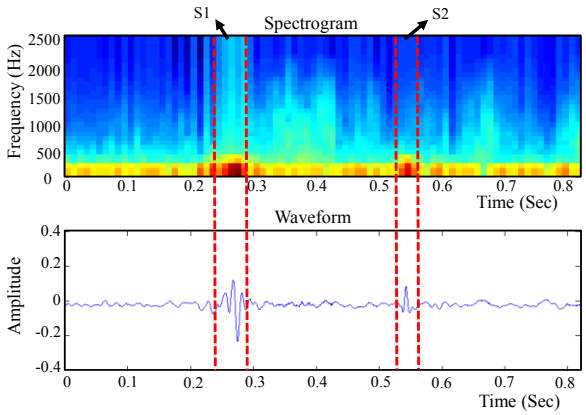


Fig. 5. Spectrogram and waveform plots of S1 and S2.

### 1) Determining optimal feature configuration and NN structure

In this section, we use the development set to determine the optimal feature configuration and NN structure. We first conducted experiments using various sound signal sampling rates: 5, 10, and 15 kHz. In this set of experiments, sound signals from both the training and development sets were processed using the same sampling rates (5, 10, and 15 kHz); a one-layer ANN model comprised of 100 hidden neurons was used as the classifier, and 13 dimensional MFCCs served as the acoustic features.

Table II lists the results of this set of experiments, indicating that the 5-kHz sampling rate sufficiently represents heart sound signals and demonstrates higher accuracy compared with the 10 and 15-kHz sampling rates. We noted that most of the heart sound signals are adequately represented at the 5-kHz sampling rate, and that the 10 and 15-kHz sampling rates include irrelevant sound components. Therefore, the 5-kHz sampling frequency provides optimal classification performance. Moreover, the signals obtained at the 5-kHz sampling rate require lower computational complexity compared with those obtained at the 10 and 15-kHz sampling rates. We maintained the sampling rate at 5 kHz in the subsequent experiments.

TABLE II
ACCURACIES OF SINGLE-LAYER ANN WITH DIFFERENT SAMPLING RATES. HIGHEST ACCURACY IS INDICATED IN BOLD FONT.

| Sampling Rate | 5 kHz | 10 kHz | 15 kHz |
|---|---|---|---|
| Accuracy | **68.33%** | 55.83% | 65.00% |

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

Next, we explore the effectiveness of the velocity and acceleration features. We extended the original MFCCs from 13 to 26 dimensions (by appending 13 velocity features) and 39 dimensions (by appending 13 velocity and 13 acceleration features). We also investigate the effectiveness of the K-means algorithm in feature extraction. Table III lists the results of the 13, 26, and 39 MFCC dimensions obtained with and without the K-means algorithm. In this study, we used the Euclidean distance for $d(\cdot,\cdot)$ in (3) for the K-means algorithm. In addition to MFCCs, we tested performance using another acoustic feature—Fbank. Unlike the MFCC features, we used 24 dimensional Fbank features, which are popularly used in DNN-based ASR systems [40]-[42]. In addition, we combined a context-window of $T$ frames around the target frame to form a concatenated Fbank feature, which incorporates the more temporal information. Accordingly, in addition to the 24 dimensional Fbank feature, we tested recognition using 168 ($24 \times 7$, $T = 3$) and 264 ($24 \times 11$, $T = 5$) dimensional Fbank features. In this set of experiments, a one-layer NN model with 100 hidden neurons served as the classifier.

TABLE III
ACCURACIES OF SINGLE-LAYER ANN USING MFCC AND FBANK WITH DIFFERENT FEATURE DIMENSIONS AND WITH/WITHOUT K-MEANS PROCESS. HIGHEST ACCURACY IS INDICATED IN BOLD FONT; (a) MFCC (b) FBANK.

(a)

| Dimension | 13 | 26 | 39 |
|---|---|---|---|
| Without K-means | 68.33% | 76.67% | 82.50% |
| With K-means | 69.17% | 79.17% | **85.00%** |

(b)

| Dimension | 24 | 168 (24×7) | 264 (24×11) |
|---|---|---|---|
| Without K-means | 81.67% | 77.50% | 76.67% |
| With K-means | 83.33% | 79.17% | 79.17% |

Table III-(a) indicates that the 39-dimensional MFCCs outperform the 13 and 26-dimensional MFCCs, thus confirming the effectiveness of using velocity and acceleration features in heart sound classification. Moreover, 39-dimensional MFCCs achieve higher recognition performance than the Fbank features listed in Table III-(b), thus indicating that the 39-dimensional MFCC feature is more suitable for this particular task. Furthermore, Table III reveals that the results obtained using the K-means algorithm are consistently superior to those obtained without using the K-means algorithm for all dimensions, verifying the effectiveness of applying the K-means algorithm to the feature extraction process. In the subsequent experiments, we applied the 39-dimensional MFCCs and K-means processing.

To further analyze the results of K-means clustering, we plotted the clustering results using 39 MFCCs and the Euclidian distance, as shown in Fig. 6. In the figure, ▲ and ■ denote the two groups clustered by the K-means algorithm. From the figure, we can observe the characteristics of the two groups: one contains the features located in the central part, and the other contains the features in the outer part of the entire heart sound segment.

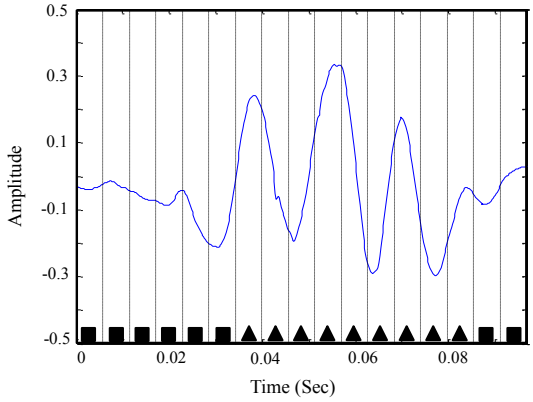


Fig. 6. K-means clustering results using 39 MFCC with Euclidian distance on S1 heart sound segment.

Next, we investigate the correlation between classification performance and NN structure (numbers of hidden layers and neurons in each hidden layer). We tested the performance of NN structures that involve one, two, and three hidden layers, with each hidden layer containing 100, 200, and 300 neurons. Table IV lists the results. Here, we can see that increasing the number of neurons from 100 to 200 and 300 does not significantly enhance the performance of the NN structures that involve one, two, and three hidden layers, suggesting that using a larger number of neurons does not lead to considerable performance enhancements in our task. However, the performance of the NN structure that involves three hidden layers is clearly superior to that involving one and two hidden layers, thus suggesting that using a deep structure can lead to higher accuracies. As indicated in Table IV, the DNN structure that comprises three hidden layers, with each layer containing 100 neurons, achieves optimal performance; this DNN structure was used in the subsequent experiments.

TABLE IV
ACCURACIES OF DNN WITH DIFFERENT NUMBERS OF LAYERS AND NEURONS. HIGHEST ACCURACY IS INDICATED IN BOLD FONT.

| Layers \ Neurons | 100 | 200 | 300 |
|---|---|---|---|
| 1 | 85.00% | 85.00% | 85.00% |
| 2 | 82.50% | 83.33% | 84.17% |
| 3 | **88.33%** | 87.50% | 83.33% |

We further tested the effectiveness of pre-training and activation functions, such as sigmoid (Sigm), hyperbolic tangent (Tanh), and rectified linear unit (Relu) [36] on the S1 and S2 classification task. In this study, we followed the pre-training and backpropagation processes presented in Section IV, and the results of with/without pre-training and different activation functions are demonstrated in Table V. Here, we first note that the pre-training procedure indeed provides improved recogni-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

tion performance when compared with that of without pre-training. Moreover, the sigmoid activation function outperforms the other two activation functions in this task.

TABLE V
ACCURACIES OF DNN WITH DIFFERENT ACTIVATION FUNCTIONS AND WITH/WITHOUT PRE-TRAINING PROCESS. HIGHEST ACCURACY IS INDICATED IN BOLD FONT.

| | Sigm | Tanh | Relu |
|---|---|---|---|
| Without Pre-training | 88.33% | 83.33% | 85.00% |
| With Pre-training | **89.17%** | 85.00% | 85.83% |

Finally, we examine the weight decay and sparsity penalty ($R(\boldsymbol{W})$ and $\rho(\boldsymbol{\mathcal{A}})$ in Eq. (7), respectively). We conducted experiments using pre-training and sigmoid activation function (the best setup in Table V) to test weight decay and sparsity penalty parameters. The experiment results are presented in Table VI, where {weight decay; sparsity penalty} = {0.0; 0.0} is the same as that presented in Table V (89.17%). As indicated in Table VI, when compared with {weight decay; sparsity penalty} = {0.0; 0.0}, {weight decay; sparsity penalty} = {0.0001; 0.01} achieves a clear accuracy improvement of 1.66% (89.17% to 90.83%). The result confirms that when the training data is limited, suitable weight decay and sparsity penalty can allow DNN attain higher S1 and S2 recognition accuracies.

TABLE VI
ACCURACIES OF DNN WITH DIFFERENT WEIGHT DECAY AND SPARSITY PENALTY SETUP. HIGHEST ACCURACY IS INDICATED IN BOLD FONT.

| Sparsity \ Weight | 0.0 | 0.0001 | 0.0005 | 0.001 |
|---|---|---|---|---|
| 0.0 | 89.17% | 86.67% | 90.00% | 89.17% |
| 0.001 | 89.17% | 88.33% | 90.00% | 90.00% |
| 0.01 | 87.50% | **90.83%** | 88.33% | 89.17% |

*2) Comparison of DNN with other classifiers*

In this section, we compare the performance of DNN with other classifiers using the testing set. As mentioned in Section V-A, S1 and S2 heart sounds that belong to the testing set were not segmented in advance. Therefore, the HSAD procedure shown in Fig. 1 and based on Shannon energy [43] was applied to detect heart sound segments before performing classification. The Shannon energy is computed by

$$\text{Shannon energy} = -x^2 \cdot \log x^2 \tag{17}$$

where $x$ is the signal.

Then, the average Shannon energy is calculated by

$$E_S = -\frac{1}{M_s}\sum_{i=1}^{M_s} x_{\text{norm}}^2(i) \cdot \log x_{\text{norm}}^2(i) \tag{18}$$

where $x_{\text{norm}}$ is the normalized signal and $M_s$ is the signal length. We set $M_s = 50$ (i.e., 10 ms) in the experiments.

From the 87 S1 and 87 S2 heart sounds in the testing set, 87 S1 and 82 S2 heart sounds were detected correctly based on the segmentation algorithm with Shannon energy [43]. Because the main purpose of this study is to investigate whether S1 and S2 can be effectively recognized based on acoustic characteristics, we tested recognition using the correctly detected 87 S1 and 82 S2 heart sounds in the following experiments.

To confirm the effectiveness of DNN on S1 and S2 classification, we tested the performance using other well-known classifiers, including KNN, GMM, LR, and SVM. The operating concept of the KNN algorithm is relatively simple because it involves using only features as the evaluation standard for distances [21]. GMM is a well-known generative model. In the training phase, a GMM is trained for each category. In the testing phase, the probabilities of testing data on GMMs of the entire categories are calculated, and the classification is determined based on probability scores [22]. The LR model is a very popular classification model that aims to maximize the conditional log-likelihood in order to optimize the model parameters [23]. The SVM classifier represents data samples as points in space and determines a gap to separate the two classes of data in the training phase. In the testing phase, each data sample is first mapped into that same space, and then predicted to belong to a class based on the side of the gap on which the sample falls [24]. In this study, the KNN classifier uses the Euclidean metric for the distance calculation. For the GMM model, eight Gaussian mixture models were used. For the SVM classifier, the Gaussian radial basis function was used as the kernel function. The parameters used in KNN, GMM, LR, and SVM were all optimized based on the development set.

Figs. 7 and 8 and Table VII demonstrate the experiment results for the recognition of S1 and S2 using KNN, GMM, LR, SVM, and DNN classifiers. The experiment results for precision, recall, $F_1$ measure, and accuracy are very consistent when comparing these five classifiers with the following three observations: (1) KNN achieves worse performance compared with the other three classifiers, possibly because of its too simple structure and limited classification capability; (2) LR gives the best S1 accuracy but the worst S2 accuracy, showing that LR might not provide balanced performance for both classes; (3) LR, SVM, and DNN outperform GMM, suggesting that discriminative classifiers can give better performance when S1 and S2 labels are given; (4) DNN achieves the best performance with 91.12% accuracy, confirming its better classification capability when compared with the other four classifiers.

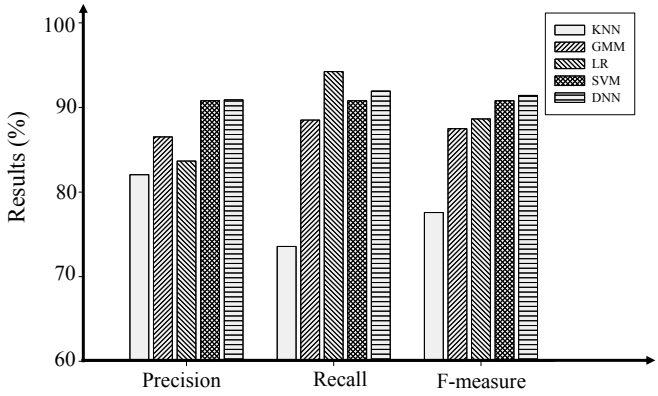


Fig. 7. Precision, Recall, and F-measure scores of S1 obtained using five types of classifiers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2559800, IEEE Transactions on Biomedical Engineering

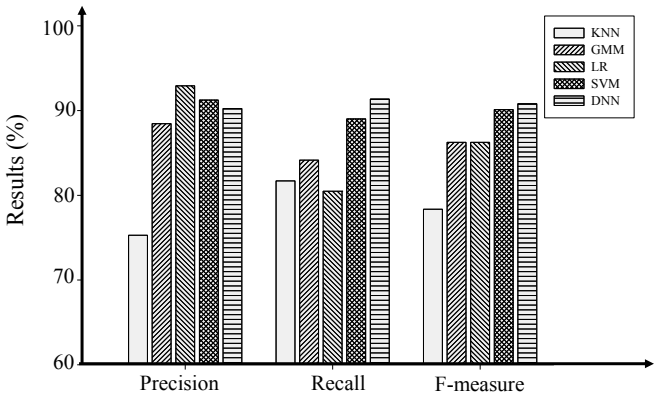


Fig. 8. Precision, Recall, and F-measure scores of S2 obtained using five types of classifiers.

TABLE VII
ACCURACIES OF FIVE CLASSIFIERS. HIGHEST ACCURACY IS INDICATED IN BOLD FONT.

| Methods | KNN | GMM | LR | SVM | DNN |
|---|---|---|---|---|---|
| Accuracy | 78.11% | 86.98% | 87.57% | 90.53% | **91.12%** |

## VI. Conclusion

This study proposed a system that entails the combination of the DNN acoustic fingerprinting classifier and MFCC acoustic features for recognizing S1 and S2 heart sounds. In this study, we focused on the recognition performed based only on the acoustic characteristics of S1 and S2 without considering the information of individual S1 and S2 duration or the time intervals for S1–S2 and S2–S1. It was noted that the recognition accuracies of individual S1 and S2 are very high, even when the duration and interval are not accessible, thus suggesting that the proposed system can be used for patients with arrhythmia, which can generate irregular S1 and S2 time intervals. The precision, recall, F-measure, and accuracy of the classification processes were compared, and the results showed that S1 and S2 can be effectively recognized with more than 91% accuracy. Furthermore, S1 recognition demonstrated better performance compared with S2 recognitions in terms of F-measure. Therefore, S1 can be used to obtain S2 according to the recognition order in clinical practice. Another contribution of this paper is that we confirmed the effectiveness of using DNN for building acoustic models to characterize S1 and S2 heart sounds. In addition, we evaluated and compared the performance of several feature extraction methods and advanced deep learning algorithms on this particular task. In the future, new corpora can be collected for experiments to increase the effect of changes in training corpora and conduct further processing in various audio-receiving environments.

## References

[1] J. Constant, Essentials of Bedside Cardiology, second edition. Humana Press Inc, 2003.

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

[3] Y. N. Wen, A. P. Lee, F. Fang, C. N. Jin and C. M. Yu, "Beyond auscultation: Acoustic cardiography in clinical practice," International journal of cardiology, vol. 172, pp. 548-560, 2014.

[4] E. G. Dimond and A. Benchimol, "Phonocardiography", California Medicine, 94(3), pp. 139–146, 1961.

[5] M. B. Malarvili, I. Kamarulafizam, S. Hussain and D.Helmi, "Heart sound segmentation algorithm based on instantaneous energy of electrocardiogram," in Proc. CinC, pp.327-330, 2003.

[6] M. El-Segaier, O. Lilja, S. Lukkarinen, L. Srnmo, R. Sepponen and E. Pesonen, "Computer-based detection and analysis of heart sound murmur," Annals of Biomedical Engineering, vol. 33, pp. 937-942, 2005.

[7] P. Carvalho, P. Gil, J. Henriques, M. Antunes and L. Eugenio, "Low complexity algorithm for heart sound segmentation using the variance fractal dimension," in Proc. ISP, pp. 194-199, 2005.

[8] H. Liang, S.Lukkarinen and I.Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in Proc. CinC, pp. 7-10, 1997.

[9] D. Kumar, P.Carvalho, M.Antunes, J. Henriques, L. Eugenio, R. Schmidt and J. Habetha, "Detection of S1 and S2 heart sounds by high frequency signatures," in Proc. EMBS, pp. 1410-1416, 2006.

[10] J. E. Hebden and J. N. Torry, "Neural network and conventional classifiers to distinguish between first and second heart sounds," IEE Colloquium, Proceedings Artificial Intelligence Methods for Biomedical Data Processing, vol. 3, pp. 1-6, 1996.

[11] A. C. Stasis, E. N. Loukis, S. A. Pavlopoulos and D. Koutsouris, "Using decision tree algorithm as a basis for a heart sound diagnosis decision support system," in Proc. EMBS, pp. 354-357, 2003.

[12] T. Olmez and Z. Dukar, "Classification of heart sounds using an artificial neural network," Pattern Recognition Letters, vol. 24, pp.617-629, 2003.

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

[14] L. Rabiner and B.-H. Juang, Fundamentals of speech recognition, Prentice Hall, 1993.

[15] C. Bishop, Pattern recognition and machine learning. Springer, 2006.

[16] J. L. Elman, "Learning and development in neural networks: the importance of starting small," Cognition, vol. 48, pp. 71-99, 1993.

[17] H. Larochelle, Y. Benqio, J. Louradour and P. Lamblin, "Exploring strategies for training deep neural networks," Machine Learning, vol. 10, pp. 1-40, 2009.

[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE, Signal Processing Magazine, vol. 29, pp. 82-97, 2012.

[19] Y. Bengio, "Learning deep architectures for AI," Foundation and Trends in Machine Learning, vol. 2, pp. 1-127, 2009.

[20] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 14–22, 2012.
[21] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, pp. 175-185, 1992.
[22] S. Theodoridis and K. Koutroumbas, Pattern recognition. Elsevier, 2009.
[23] E. Frank and J. Harrell, Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis, Springer, 2001.
[24] C. Campbell and Y. Ying, Learning with support vector machines. Morgan & Claypool, 2011.
[25] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," in Proc. ICASSP, vol. 28, pp.357-366, 1980.
[26] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustic, Speech, Signal Processing, vol. ASSP-29, pp. 254–272, 1981.
[27] B. Logan, "Mel frequency cepstral coefficients for music modeling," in Proc. ISMIR. 2000.
[28] S. Chauhan, P. Wang, C. S. Lim and V. Anantharaman, "A computer-aided MFCC-based HMM system for automatic auscultation," Computers in Biology and Medicine, vol. 38, pp. 221-233, 2008.
[29] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 23, pp. 7-19, 2015.
[30] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in Proc. Interspeech, 2013.
[31] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in Proc. ICDAR, pp. 958–963, 2003.
[32] D. Ciresan, U. Meier and J. Schmidhuber, "Multi-column deep neural networks for image classification," in Proc. CVPR, pp. 3642-3649, 2012.
[33] G. Hinton, A practical guide to training restricted Boltzmann machines, Springer, 2011.
[34] D. Yu, L. Deng, Automatic speech recognition: A deep learning approach, Springer, 2014.
[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371-3408, 2010.
[36] X. Glorot, A. Bordes and Y. Bengio, "Deep sparse rectifier neural networks," in Proc. AISTATS, pp. 315-323, 2011.
[37] S. M. Siniscalchi, T. Sveendsen, and C.-H. Lee, "An artificial neural network approach to automatic speech processing," Neurocomputing, pp.326-338, 2014.
[38] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET curve in assessment of detection task performance," in Proc. Eurospeech, pp. 1895-1898, 1997.
[39] J. L. Flanagan, Speech analysis, synthesis and perception. Springer-Verlag, 1972.
[40] D. Yu, M. L. Seltzer, J. Li, J. T. Huang and F. Seide, "Feature learning in deep neural networks- studies on speech recognition tasks," 2013, http://arxiv.org/pdf/1301.3605.
[41] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 30-42, 2012.
[42] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Speech recognition using long-span temporal patterns in a deep network model," IEEE Signal Processing Letters, vol. 20, pp. 201-204, 2013.
[43] H. Liang, S. Lukkarinen and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in Proc. CinC, pp. 105-108, 1997.