Journal of Image and Signal Processing 图像与信号处理, 2018, 7(4), 213-219
Published Online October 2018 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2018.74024

Hans汉斯

Principal Component Analysis-Based Heart Sound Features Reduction Research

Junjie Zhang, Biqiang Zhang, Dingbin Wang, Feiyan Hu

Department of Electronics and Electrical Engineering, Nanyang Institute of Technology, Nanyang Henan
Email: mr_right523@yeah.net

Received: Sep. 25th, 2018; accepted: Oct. 10th, 2018; published: Oct. 17th, 2018

Abstract

In this research, a method based on principal component analysis (PCA) is proposed for using the optimal dimensional features to describe the distribution of heart sound characteristics in different kinds of heart diseases. This study is described in three stages. In stage 1, heart sound signal is collected via 3M-3200 electronic stethoscope and preprocessed based on wavelet transform. In stage 2, the power spectrum density combined with threshold method is proposed to define the cardiac sound 7-dimensional feature. In stage 3, based on principal component selection criteria combined with scatter plot distribution results, the final heart sound is determined to be a 2-dimensional feature representing 96.1% information of 7-dimensional feature. The results of the research on the typical heart diseases indicate that there are obviously differences in the distribution of the heart sound features among different kinds of heart diseases.

Keywords

Principal Component Analysis (PCA), Wavelet Decomposition, Power Spectrum Density

基于主成分分析的心音特征降维处理研究

张俊杰，张弼强，王丁彬，胡飞燕

南阳理工学院电子与电气工程学院，河南 南阳
Email: mr_right523@yeah.net

收稿日期：2018年9月25日；录用日期：2018年10月10日；发布日期：2018年10月17日

摘 要

本研究提出一种基于主成分分析的心音特征降维处理方法，实现以最优维度描述不同种类心脏病的心音

文章引用: 张俊杰, 张弼强, 王丁彬, 胡飞燕. 基于主成分分析的心音特征降维处理研究[J]. 图像与信号处理, 2018, 7(4): 213-219. DOI: 10.12677/jisp.2018.74024

特征分布。本文分三阶段进行论述：第一阶段，基于美国3M公司3200型电子听诊器的心音信号采集及基于小波变换的心音预处理；第二阶段，利用功率谱对心音信号进行频域分析，采用阈值法定义心音频域特征；第三阶段，基于主成分选取准则并结合散点图分布结果，确定以表征七维特征96.1%信息量的两维特征作为最终心音量。典型心脏病例的研究结果表明，异类心脏病心音特征分布呈现出明显区分。

关键词

主成分分析，小波分解，功率谱

1. 引言

根据《中国心血管病报告 2017(概要)》[1]，我国 2016 年农村和城市居民主要疾病死因构成比如图 1 所示，其表明农村和城市心血管疾病死亡人数均已超过疾病死亡人数的 40% [1]。因此心血管病已成为危害我们健康的重疾之首，其预防和治疗至关重要，迫在眉睫。

心脏病的及时正确诊断是心脏病治愈的前提和基础，而心音分析是诊断心脏病的一种重要途径，其具有费用低廉、无创伤性、简便有效等优点，被广泛用于心脏疾病的诊断分析。而利用心音分析诊断心脏病中起决定性的中间环节是心音特征提取，而心音特征数据维数的不断增长已严重影响诊断效率，在很多聚类问题中(如机器学习[2]、图像处理[3]、模式识别[4]、文本分析[5]等)特征降维是处理高维数据的一个重要步骤。降维处理方法主要有三种方法：Filter 方法、因子分析方法和主成分分析方法(PCA)。其中，PCA 算法简单有效且无参数限制，因此被广泛应用于数据压缩和特征提取。鉴于此，本研究提出一种基于主成分分析的心音频域特征降维处理方法，实现以最优维度表征心音频域特征、以可视化效果表征心音频率分布。为验证本研究提出方法的有效性，以常见的典型心脏病例作为研究对象，其特征分布结果表明：不仅特征分布可视化效果明显，而且不同类心音特征呈现出显著区分。因此本研究提出的方法可为医护人员及研究人员提供一种较为明确的可视化诊断信息。

2. 心音信号的采集及预处理

2.1. 心音信号的采集

在临床上，心脏瓣膜听诊区通常有四个：二尖瓣区、肺动脉瓣区、主动脉瓣区、三尖瓣区[2]。本研究采用美国 3M 公司生产的 3200 型电子听诊器[3]在主动脉瓣听诊区进行心音采集(采样频率 $F_s = 44.1\,\text{kHz}$)，采集心音实例图及听诊器实物图如图 2 所示。

2.2. 心音信号的预处理

心音信号极其微弱，在采集的过程中极易受到噪音的干扰，造成部分有用信息的丢失及心音的识别度降低。此外，心音信号的复杂性和非平稳性使得对其进行分析变得困难，再加上噪声的引入进一步增加了分析心音信号的难度。因此，在采集过程中应尽量避免噪声干扰和对心音进行降噪处理是必要的。研究[4] [5]表明在基于小波分解的预处理中，MATLAB 函数波 *wavedec* 和 *waverec* 是根据心脏的活动特征来实现的。基于采样频率 F_s 中所描述的声音频率范围，10 层近似系数(0~21.5 Hz)用于切断低于 21.5 Hz

的低频成分，而第 5 层近似系数(689~1387 Hz)用于消除 689 Hz 的高频成分。过滤后的信号与 21.5~689 Hz 组件的频带限制是由 6 层到 10 层的细节系数组成的。由于 DB10 小波给出了最大的信噪比(SNR)，所以本研究采用 DB10 作为母小波对心音进行降噪处理，对二尖瓣回流心音信号的降噪处理如图 3 所示。由图可知，DB10 可以有效地去除心音中的噪声，同时保留信号所携带的有用信息。

3. 心音信号的频域特征提取

研究表明，不同种类的心音信号具有不同的频率分布，鉴于此，本文提出一种利用阈值的方法通过心音的包络线提取不同阈值对应的不同频率成分进行分析如下：第一步在频率域内提取心音包络公式如公式(1)至(3)。第二步利用阈值线提取心音特征如图 4 所示。图 5 为不同阈值所对应的不同心音(主动脉

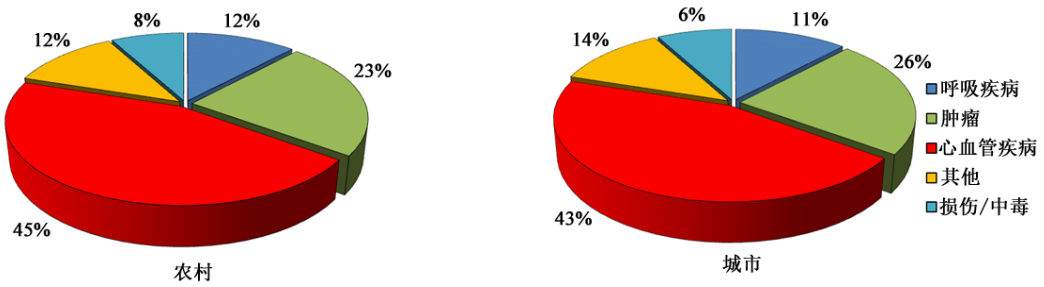

Figure 1. Proportion of death among rural and urban residents in China in 2015
图 1. 2015 年中国农村和城市居民主要疾病死因构成比

Figure 2. Collection of heart sounds examples and stethoscope physical chart
图 2. 采集心音实例图及听诊器实物图

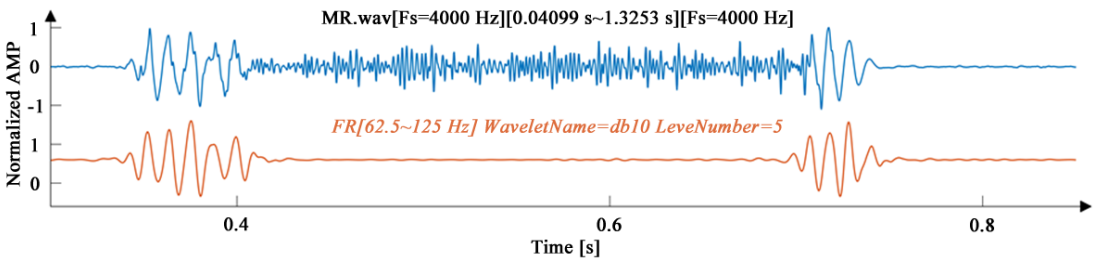

Figure 3. An example of noise reduction of mitral regurgitation signal
图 3. 二尖瓣回流心音信号的降噪示例图

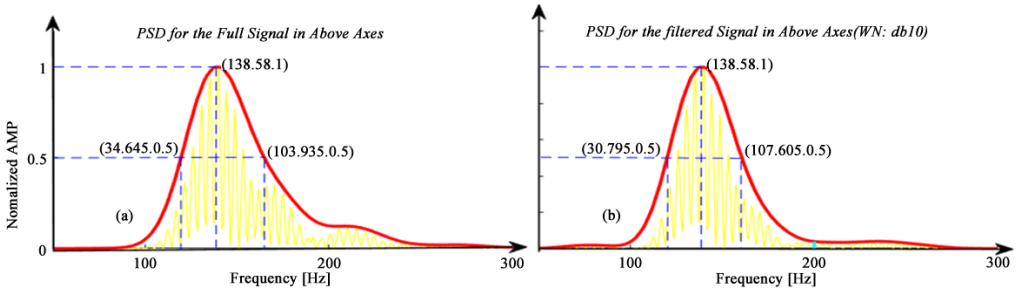

Figure 4. Frequency domain feature extraction
图 4. 频域心音特征提取定义示意图

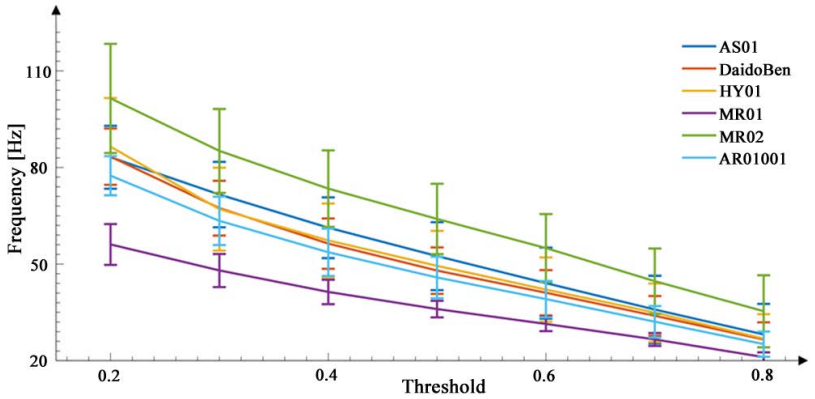

Figure 5. Characteristic statistical chart ($\mu \pm \sigma$)
图 5. 特征统计图($\mu \pm \sigma$)

狭窄(AS)、二尖瓣回流(MR)、主动脉回流(AR)和正常心音(NM))特征统计图($\mu \pm \sigma$)。从图中可以看出，各类心音特征之间还是有一定差别的，但由于特征维数较高，无法确定用哪几个特征能够进行心音的可视化分类识别。因此，下一节详述基于主成分分析的心音特征降维处理及结果。

包络线 E_{F_i}，第i个心音周期$\left(X_{T_{Pi}}\right)$，对于诊断心室间隔缺损来说，已经被证实有一种有效的提取频域特性的方法[6]。使用相同的方法，依据公式(1)和(2)提取包络线 E_{sj}：

$$E_{sj}[k]=\frac{1}{2L_{Sj}+1}\sum_{l=k-L_{Sj}}^{k+L_{Sj}}\left|X_{F_{Sj}}[l]\right|,k=L_{Sj},\cdots,N_{Sj}-1-L_{Sj},j=1,2,3 \tag{1}$$

$$X_{F_{Sj}}[l]=\sum_{n=1}^{N_{Sj}}X_{T_{Sj}}[n]\mathrm{e}^{\left(-j\frac{2\pi}{N_{Sj}nl}\right)}l=0,1,2,\cdots,N_{Sj}-1 \tag{2}$$

其中 $X_{F_{Sj}}$ 为心音周期，$X_{T_{Sj}}$ 为提取的第 i 个心音周期，N_{Sj} 为第 i 个心音周期的长度。

为了使包络面有利于分析频率成分，在研究中，基于包络线 E_{sj}，提出二次包络线 SE_{Sj} 的公式如下：

$$SE_{Sj}[k]=\frac{1}{2SL_{Sj}+1}\sum_{l=k-SL_{Sj}}^{k+SL_{Sj}}\left|E_{Sj}[l]\right|,k=SL_{Sj},\cdots,N_{Sj}-1-SL_{Sj},j=1,2,3 \tag{3}$$

其中 $2SL_{Sj}=35\times\left(F_S\right)/\left(N_{Sj}\right)$ 是为第 i 个心音信号设置的窗宽。

4. 降维处理研究

4.1. 主成分分析的简介

主成分分析(PCA)是一种对高维数据进行线性降维的统计方法，广泛应用于心脏病[7]，甲状腺疾病[8]，

心脏死亡[9]，冠状动脉疾病[10]，数据聚类[11]，心血管疾病[12]和心律不齐[13]等研究中。

主成分分析的具体步骤如下：

1) 对原始数据进行标准化处理(消除量纲影响)

设特征矩阵为 $X=\begin{bmatrix} \chi_{11} & \cdots & \chi_{1m} \\ \vdots & \ddots & \vdots \\ \chi_{n1} & \cdots & \chi_{nm} \end{bmatrix}$ 则按公式(4)进行标准化处理：

$$\chi_{ij}^{*}=\frac{\chi_{ij}-\overline{\chi}_{j}}{\sqrt{Var\left(\chi_{j}\right)}} \tag{4}$$

其中，$\overline{\chi}_{j}=\frac{1}{n}\sum_{i=1}^{n}\chi_{ij}$，$Var\left(\chi_{j}\right)=\frac{1}{n-1}\sum_{i=1}^{n}\left(\chi_{ij}-\overline{\chi}_{j}\right)^{2}$，$i=1,2,3,\cdots,n$，$j=1,2,3,\cdots,m$。

2) 计算数据的协方差矩阵

协方差公式(5)计算：

$$X=\left(s_{ij}\right) \tag{5}$$

3) 求出矩阵的特征值及相应的正交化单位特征向量

解方程$\left|\lambda E-\mathrm{X}\right|=0$，求出特征值$\lambda_{i}$，按从大到小的顺序排列$\lambda_{1}\geq\lambda_{2}\geq\cdots\geq\lambda_{7}$，并求出$\lambda_{i}$对应的特征向量$a_{i}\left(i\leq 7\right)$。

4) 计算各特征值的贡献率及累计贡献率

贡献率及累积贡献率分别依据公式(6) (7)计算：

$$\frac{\lambda_{i}}{\sum_{k=1}^{m}\lambda_{k}} \tag{6}$$

$$\frac{\sum_{k=1}^{i}\lambda_{k}}{\sum_{k=1}^{m}\lambda_{k}} \tag{7}$$

4.2. 特征降维处理结果

根据以上步骤在 MATLAB 中运行主成分分析程序，得到的矩阵的主成分系数如表 1 所示，其对应的特征值及其累计贡献率如表 2 所示，随着特征值的逐渐减小，累计贡献率趋近于 1；当特征值较小时，该成分贡献率接近于 0，因此表 2 中后三个特征值不进行考虑，得出特征降维处理的结果如图 6 所示。

Table 1. Principal component analysis
表 1. 主成分系数

特征	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	0.4890	0.7652	0.3766	−0.1809	−0.0150	0.0239	−0.0003
2	0.4588	0.0849	−0.4521	0.5842	0.3469	−0.3469	−0.0102
3	0.4125	−0.1069	−0.3721	0.0095	−0.3317	0.7082	−0.2616
4	0.3759	−0.2311	−0.1744	−0.3288	0.3046	−0.2155	0.7263
5	0.3333	−0.3239	0.0360	−0.4515	−0.0161	−0.4516	−0.6121
6	0.2754	−0.3636	0.3277	−0.1237	0.7161	0.3570	0.1706
7	0.2282	−0.3245	0.6134	0.5465	−0.4045	−0.0618	−0.0147

Table 2. Eigenvalue and its cumulative contribution rate
表 2. 特征值及其累积贡献率

特征值	903.3046	87.8791	24.8642	5.8831	1.2395	0.9171	0.2484
累积贡献率	0.8819	0.9677	0.9919	0.9977	0.9989	0.9998	1.0000

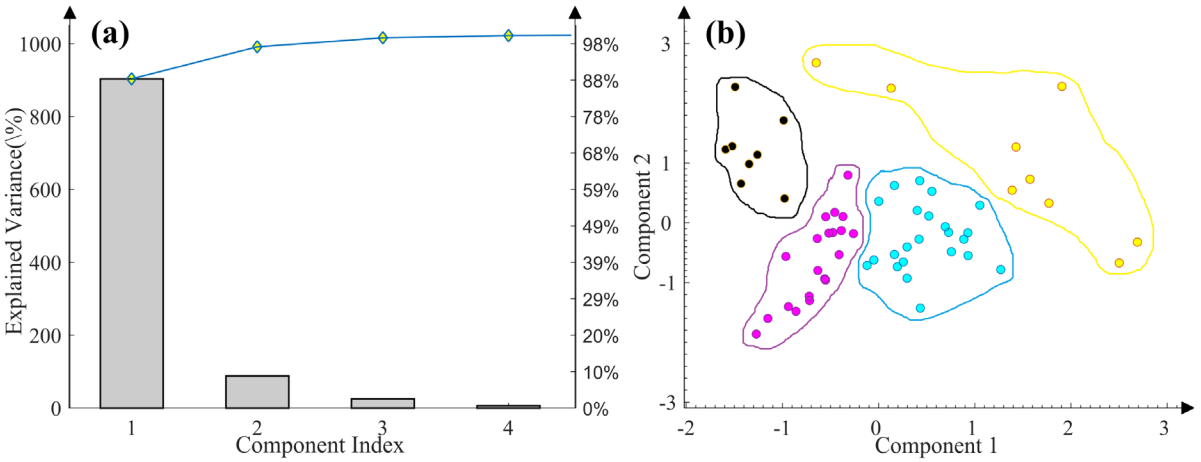

Figure 6. Principal component analysis results
图 6. 主成分分析结果

5. 结论

本研究针对频域中心音的特征分布，针对频域中心音的特征分布，本研究提出一种基于主成分分析的心音频域特征降维处理方法，采用帕累托图、散点图并结合主成分选取准则将特征从 7 维降到 2 维，实现了 7 维特征数据的 96.1%的信息量实现以最优维度表征心音频域特征、以可视化效果表征心音频率分布。不同种类心脏病心音信号特征分布散点图表明本文提出的特征提取方法能够效地表征心音信号的特征，后续还需要采用模式识别算法对心音进行分类识别分析。

参考文献

[1] 《中国心血管病报告 2017》概要[J]. 中国循环杂志, 2018(1): 1-8.

[2] 罗玉光, 心脏听诊的基础知识——怎样听心音[J]. 人民军医, 1964(7): 41-44.

[3] https://www.littmann.com

[4] 陈新华, 成谢锋. 一种改进型综合去噪算法在心音信号预处理上的研究[J]. 南京邮电大学学报(自然科学版), 2010, 30(6): 96-100.

[5] 单正娅. 基于人工神经网络及小波分析的心音诊断系统的研究[D]: [硕士学位论文]. 无锡: 江苏大学, 2006.

[6] Sun, S.P., *et al.* (2014) Segmentation-Based Heart Sound Feature Extraction Combined with Classifier Models for a VSD Diagnosis System. *Expert Systems with Applications*, **41**, 1769-1780. https://doi.org/10.1016/j.eswa.2013.08.076

[7] 刘立汉, 王海滨, 王燕, 陶婷, 魏秀波. 基于改进的希尔伯特－黄变换的心音信号特征分析[J]. 西华大学学报(自然科学版), 2010, 29(6): 14-18.

[8] 胡玉良, 王海滨, 陈健, 等. 心音在时频两域中解析方法的研究[J]. 西华大学学报(自然科学版), 2009(5): 5-8, 26.

[9] Joy, R., Acharya, U.R., Mandana, K.M., Ray, A.K. and Chakraborty, C. (2012) Expert Systems with Applications Application of Principal Component Analysis to ECG Signals for Automated Diagnosis of Cardiac Health. *Expert Systems with Applications*, **39**, 11792-11800. https://doi.org/10.1016/j.eswa.2012.04.072

[10] Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T.-C., Ahamed, T. and Suri, J.S. (2013) Automated Diagnosis of Coronary Artery Disease Affected Patients Using LDA, PCA, ICA and Discrete Wavelet Transform. *Knowledge-Based*

Systems, **37**, 274-282. https://doi.org/10.1016/j.knosys.2012.08.011

[11] Lee, J. and Jun, C.-H. (2013) PCA-Based High-Dimensional Noisy Data Clustering via Control of Decision Errors. *Knowledge-Based Systems*, **37**, 338-345. https://doi.org/10.1016/j.knosys.2012.08.013

[12] Shilaskar, S. and Ghatol, A. (2013) Feature Selection for Medical Diagnosis: Evaluation for Cardiovascular Diseases. *Expert Systems with Applications*, **40**, 4146-4153. https://doi.org/10.1016/j.eswa.2013.01.032

[13] Zhu, B., Ding, Y. and Hao, K. (2014) Multiclass Maximum Margin Clustering via Immune Evolutionary Algorithm for Automatic Diagnosis of Electrocardiogram Arrhythmias. *Applied Mathematics and Computation*, **227**, 428-436. https://doi.org/10.1016/j.amc.2013.11.028