学号 2017301500098

密级

武汉大学本科毕业论文

基于咳嗽音音频分析的哮喘检测

院（系）名　称：计算机学院

专　业　名　称：计算机科学与技术

学　生　姓　名：张建

指　导　教　师：张健　教授

二〇二二年三月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：_____ 日期：_____

摘　　要

智能手机和可穿戴设备的人体健康检测系统已成为诊断医学和计算机应用交汇的热点。文研究主要利用音频建模技术对哮喘这一种社会常见呼吸道疾病进行诊断。文本对国内外人体信号检测系统进行充分的研究，发现移动设备健康检测作为一种新兴的监测手段，具有易测量、诊断时间短、特异性强等显著优势，有很强的实用价值。

本研究基于以上背景，实现了一种基于咳嗽音频分析哮喘检测系统。该系统利用移动设备的麦克风以 44.1Khz 的采样率收集用户的咳嗽音频，并通过咳嗽检测网络检测音频中的咳嗽事件并选择样本质量较好的咳嗽音频。接着，本系统将过滤后的咳嗽音频重采样到 22Khz，通过音频增强技术扩增咳嗽音频的数据量，并提高整体系统对复杂情况的鲁棒性。之后系统按照人工设计特征提取和 VGG 预训练模型迁移学习特征提取两种方式，对每个单周期咳嗽音频提取 733 维度的特征向量。在对提取的特征向量进行主成分分析法完成降维处理后，系统对所有的特征向量采取 RBF-SVM 进行二分分类，并验证分类效果。

在实际的检测环节，由于咳嗽检测网络通过单周期多层决策提高决策正确率，在环境噪音大的情况下，决策失败较高，往往需要多次样本收集才能完成咳嗽检测的决策。

实验阶段实验组采集了 4 名患有哮喘的病人采集了 10min 左右的咳嗽信号筛选了共 300 组单周期咳嗽信号，同时采集了健康人的 500 组单周期咳嗽信号。利用这些数据研究了音频中咳嗽事件的检测、利用迁移学习实现特征提取以及支持向量机份分类的性能问题。最终优化系统网络参数，选取了训练集-测试集比率为 80:20 的情况并采用 RBF-SVM 进行特征分类，达到了 91.61% 的分类正确率。

关键词：哮喘检测；咳嗽音频，音频分析，迁移学习，VGG 网络；机器学习，支持向量机

ABSTRACT

The human health detection system of smart phones and wearable devices has become a hot spot at the intersection of diagnostic medicine and computer applications. This research mainly uses audio modeling technology to diagnose asthma, a common respiratory disease in society. The text conducted a thorough research on human body signal detection systems at home and abroad, and found that mobile device health detection, as an emerging monitoring method, has significant advantages such as easy measurement, short diagnosis time, and strong specificity, and has strong practical value.

Based on the above background, this research has implemented an asthma detection system based on cough audio analysis. The system uses the microphone of the mobile device to collect the user's cough audio at a sampling rate of 44.1Khz, and detects cough events in the audio through the cough detection network and selects cough audio with better sample quality. Next, the system resamples the filtered cough audio to 22Khz, amplifies the data volume of the cough audio through audio enhancement technology, and improves the robustness of the overall system to complex situations. After that, the system extracts a 733-dimensional feature vector for each single-cycle cough audio according to two methods: manual design feature extraction and VGG pre-training model migration learning feature extraction. After performing the principal component analysis method on the extracted feature vectors to complete the dimensionality reduction processing, the system adopts RBF-SVM for binary classification of all feature vectors, and verifies the classification effect.

In the actual detection link, because the cough detection network improves the accuracy of decision-making through single-cycle multi-layer decision-making, decision-making failures are high in the case of large environmental noise, and it often requires multiple sample collections to complete the cough detection decision-making.

In the experimental stage, the experimental group collected 4 patients with asthma, collected cough signals for about 10 minutes, screened a total of 300 groups of single-cycle cough signals, and collected 500 groups of single-cycle cough signals from healthy people. Using these data, the performance problems of detecting cough events in audio,

using transfer learning to achieve feature extraction, and support vector machine classification are studied. Finally, the system network parameters were optimized, the training set-test set ratio was selected as 80:20, and RBF-SVM was used for feature classification, achieving a classification accuracy rate of 91.61%.

Key words: Asthma Detection; Cough audio; Audio Analysis; Transfer Learning; VGG Network; Machine Learning; Support Vector Machine

目　　录

图片索引

表格索引

1 绪论

1.1 引言

哮喘是现代社会常见一种呼吸道疾病，患者常有喘息、咳嗽、呼吸短促等症状。因为其具有高发病率，持续时间长和反复发作等特性[1]，已经成为了严重的社会公共卫生问题。同时在所有的患者中只有 25%~50% 患者被医生知晓，众多的患者在早期未被诊断前会出现肺功能逐步下降等病理现象，这对于患者的恢复与治疗产生了更大的挑战。[2]

长期以来，医务人员已经认识到人体声音可以作为健康的指标。例如，使用听诊器听取来自心脏或肺部的声音[3]。但是这些通常需要熟练的临床医生进行聆听和诊断，并近期迅速被各种成像技术所替代，诸如 MRI、超声检查等。而对于这些新兴技术而言，分析和诊断疾病更加容易，但是并不适用于大规模的潜在患者的筛选，不具有经济性和便捷性。用自动音频特征建模技术将会给个人的日常生活工作的疾病监测提供巨大的潜力。

同时智能手机和可穿戴设备已经被广泛的使用于人体信号监测，例如：呼吸波形[4]、脉搏[5]、肌肉振动[6] 等身体信号。例如：在 breathListener[4] 中，手机麦克风和扬声器生成 ESD 信号对人体胸腔的运动变化进行监测，生成了精度高的人体呼吸波形。但是以上研究仅仅局限于提高数据收集精度，而并没有用于实际的医学诊断和治疗中，作为一种有潜力且最方便的信号监测工具，智能手机可以作为一般用户自我诊疗的辅助系统，加强普遍场合下健康诊断的可靠性。

为了解决以上痛点，本论文实现了一种基于音频特征建模技术的哮喘检测方法。对比与传统的哮喘检测方法，音频特征建模技术具有易于测量、诊断时间短、特异性强等优势：

1. **易于测量**：一般来说，咳嗽音的频率低于 20kHz, 而一般的智能手机麦克风的采样频率在 40kHz 左右。因此，本研究可以直接应用智能手机的麦克风进行咳嗽音的采样，并将咳嗽音重采样至 22kHz, 极大的拓展了哮喘检测的使用场景和应用范围。
2. **诊断时间短**：对于已经生成的分类模型，将测试好的音频输入模型进行处理、分类，可以在 2 分钟内返回诊断结果。

3. **特异性强**：前期研究实验表明，即使咳嗽不是自发的，也就是哮喘患者被要求咳嗽时，它仍然含有哮喘的特定特征[7]。这意味着受试者可以通过模拟咳嗽的方法来对哮喘进行筛选检测，并且具有一定的稳定性与特异性。

1.2 本论文的主要工作

本论文实现了一种基于音频特征建模技术的哮喘检测方法，其主要工作和贡献概述如下：

1. **采样应用实现**：基于 Android Studio 平台，实现了一款名为 asthma sound 的安卓应用，采样 3 次咳嗽的录音存储到手机本地内存。
2. **音频数据增强**：由于采样数据有限，系统在检测采样音频为咳嗽音后，将采样音频重新定位到 22kHz, 采用原始信号放大，添加白噪音，更改音调和速度共三种方法增强音频[8]。在合理的范围内将每种方法分别应用 2 次，将数据量扩增 6 倍，用于原始模型的训练，从而大大提高模型的鲁棒性。
3. **音频特征提取**：基于 librosa[9] 音频处理库，利用重采样的音频在帧和段层次使用 MFCC 提取各种音频特征，同时利用 VGG 网络自动提取其他维度的音频特征，共提取了 733 维的特征向量，最后采用主成分分析（PCA）法进行降维，用于音频分类。
4. **特征分类网络实现**：对于所有提取到的特征向量，系统使用支持向量机（SVM）进行二分类，并且使用 K 折交叉验证，减少数据的过拟合，分类出哮喘与健康人群。
5. **分类模型评估**：对于本模型实现的二分类模型，统计所有的特征数据的分类结果，并统计为混淆矩阵进行分析。混淆矩阵共有四个分量：TP（实际为正预测为正），FP（实际为负但预测为正），TN（实际为负预测为负），FN（实际为正但预测为负）。具体如图1.1所示。

Confusion Matrix		Predict		
		0	1	Total
Actual	0	TN(True Negative)	FP(False Positive)	TN+FP(Actual Positive)
	1	FN(False Negative)	TP(True Positive)	TP+FN(Actual Negative)
	Total	TN+FN(Predict Positive)	TP+FP(Predict Positive)	TP+FP+TN+FN

图 1.1　模型评估指标

1.3 文本组织架构

本文基于对近期研究热点音频特征建模技术，将其与目前健康检测领域的痛点相结合，并聚焦于目前一种常见的呼吸道疾病——哮喘，提出了一种基于音频特征建模的哮喘检测系统。该论文的总体组织结构如下：

第一章为绪论。本章简单介绍了当前健康检测领域的发展现状，并分析了当前健康检测领域的痛点和音频特征建模技术的优势。最后介绍了本论文的主要工作点。

第二章为理论基础。本章首先介绍音频特征建模工作的基础，然后介绍音频数据增强方法，并分析音频增强方法对实验结果的影响、最后介绍深度卷积网络 VGG 的基本原理，并简单分析利用 VGG 提取音频特征的原因。

第三章为系统设计。本章首先介绍本论文系统的整体架构，然后分模块介绍最重要的四个部分：咳嗽音频获取与筛选、咳嗽音频的数据增强、特征提取网络的架构以及特征分类网络的架构。

第四章为实验设计及其结果。本章介绍如何利用 k-折交叉验证设计实验验证设计模型的有效性；并将结果进行分析说明

第五章为总结与展望。本章对本文的研究工作进行总结，分析了本文提出模型的优缺点。并在该课题下对本研究未来的方向提出了新的展望。

2 理论基础

2.1 音频特征建模基础

在深度学习领域，音频特征建模常用于语音识别等领域。目前市面常见的虚拟助手：Siri、小爱同学和图灵机器人等，都是构建于音频特征建模的基础上。在音频分类、语言合成和语言认识方面，音频特征建模已经很成熟了。在深度学习领域，音频特征建模常用于语音识别等领域。目前市面常见的虚拟助手：Siri、小爱同学和图灵机器人等，都是构建于音频特征建模的基础上。在音频分类、语言合成和语言认识方面，音频特征建模已经很成熟了。基于音频处理库——Librosa，本部分主要介绍音频特征建模的理论基础：数字信号处理、滤波器和 Mel 图[9]

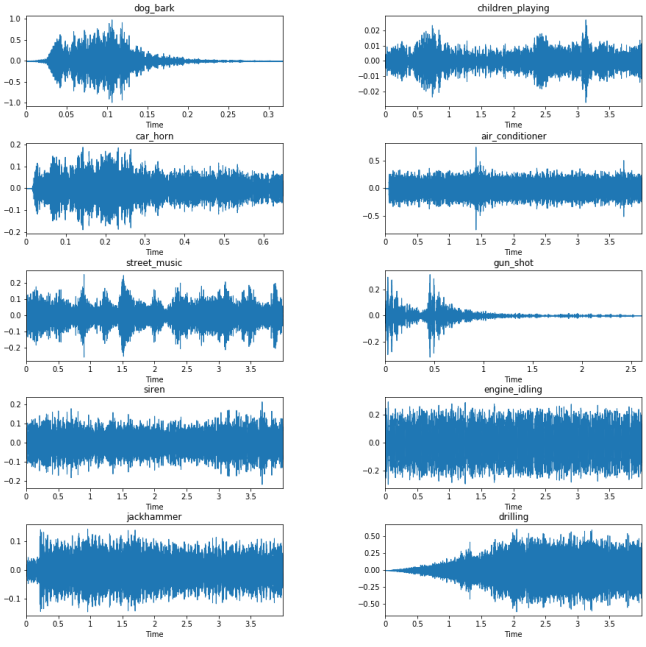


图 2.1 常见音频的波形图

2.1.1 从原始音频说起

常见的音频主要是利用手机麦克风，在 40kHz 左右的采样频率下进行收集的。因此呈现在 wav 文件中常常是以一维的格式利用扬声器进行输出。从人眼观察的

角度来看，音频样本具有着一定的周期性和特征性，但是更精细的信息都是人眼无法去分别出来的，图图2.1中有个十种音频信号的波形图，通过肉眼观察，引擎声、报警器声和空调声看上去非常的相似。以下音频信号的概念是常见用于描述波形图的特征的术语。

1. **采样和采样频率**：在音频信号处理中，采样是将连续信号按一定规律记录信号使之成为一组离散值而采样频率就是这个采样的规律即一定时间内采样的个数。一般常见手机麦克风的采样频率在 40kHz 左右。
2. **幅值**：音频波形的幅度是固定时间内波形变化的量度。幅度的另一个常用定义是变量的极差的大小。一般为了比较相对度的大小，系统会在音频处理种把幅度分量进行归一化处理。

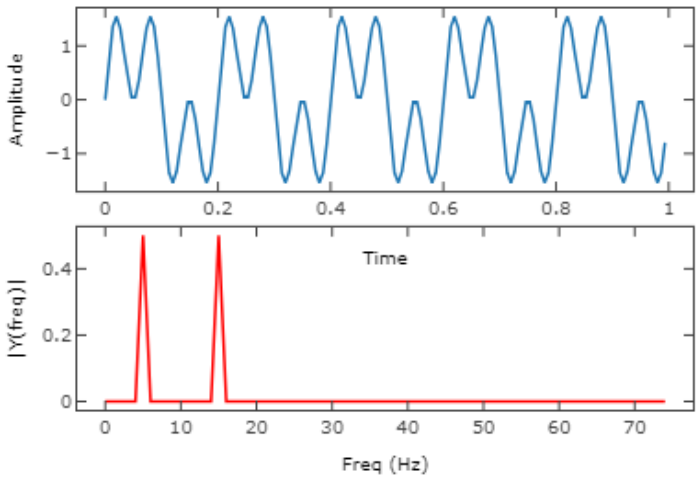


图 2.2　傅里叶变换示例

3. **傅里叶变换**：原始声音波形往往只能看到其随时间变化规律，而看不出来其频率的变化规律。傅里叶变换就是为了展示其频率变化的一种方法。

$$F(w) = \mathrm{F}[f(t)] = \int_{-\infty}^{+\infty} f(t)e^{-iwt}dx \tag{2.1}$$

傅里叶变换的理论基础是将所有信号看作任意多个三角函数的叠加。因此傅里叶变换是一种将信号投影到三角函数（基）的方法。公式（2.1）中 w 表示频率，t 表示时间，（2.1）中将信号表示为三角基函数的叠加。对于一个非周期信号，可以将非周期信号看成一个周期信号的一部分。即：傅里叶变换当周期足够大，就会退化为一般情况下的信号函数。因此傅里叶级数退化为：

$$f(t) = \sum_{k=-\infty}^{\infty} C_k e^{2\pi(K/T)t} \tag{2.2}$$

其中 C_k 为傅里叶级数，其表示为：

$$C_k = \frac{1}{T}\int_0^T e^{-2\pi i(\frac{k}{T})t} f(t)dt = \frac{1}{T}\int_{-\frac{2}{T}}^{-\frac{2}{T}} e^{-2\pi i(\frac{k}{T})t} f(t)dt \tag{2.3}$$

图2.2是利用 matlab 生成的周期音频信号。上图是一个周期音频信号，下图是该信号的变换结果。

4. **周期图**：周期图是基于傅里叶变换的方法。它对音频信号直接做傅里叶变换并且进行平方，如公式2.5：

$$S_{pre}(w) = \frac{1}{N}|c_k|^2 \tag{2.4}$$

其中 C_k 是傅里叶级数。周期图代表了一个音频在频率方向上密度的估计，即图2.3表示的是图2.2中音频信号的周期图。

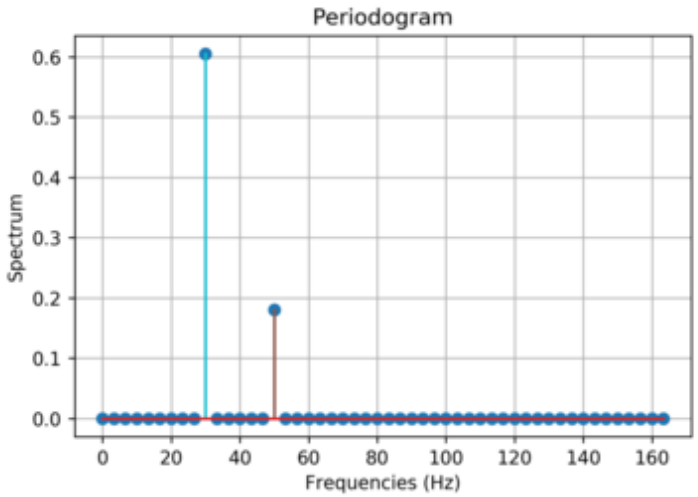


图 2.3　周期图示例

5. **梅尔倒频谱系数**：由于人耳对等距离音高感官并不是非线性的。在 1980 年 Davis 和 Mermelstei[10] 提出一种定义: 将 1000Hz 且比人类听觉阈值高 40 分贝高的音频信号定义为 1000mels。当频率大于 500Hz 时，每当人耳感觉到相同的音高变化量时，所需的频率变化就会随着频率的增加而变得越来越大。

$$m = 2595\log_{10}(1 + \frac{f}{700}) \tag{2.5}$$

图2.4是频率到 Mel 的映射关系图。从图中可以看到，在频率较低时，Mel 随频率变化较快；当频率较高时，斜率变小，变化缓慢[10]。

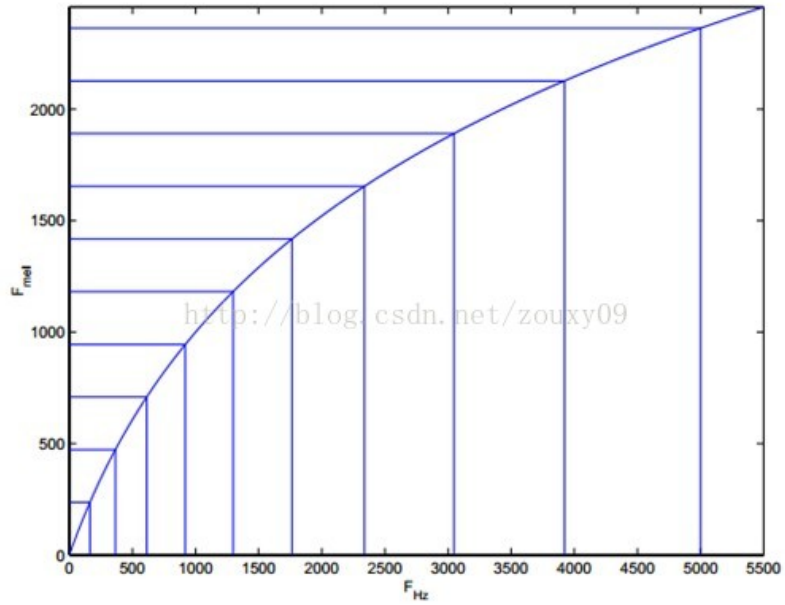


图 2.4　频率:Mel 关系示意图

2.1.2　音频特征提取之 Mel 图

音频特征提取的一个重要方面就是：梅尔频谱系数（MFCC）的计算，而利用 MFCC 可以求出 Mel 图。Mel 图与其他一般的音频特征提取方法相比共有以下优势：

1. 对于音频的特征进行了去除相关度并进行压缩，更加方便后续的处理。
2. 适合数据量更小的样本集。
3. MFCC 具有低维度，在频谱上显示的更加平整。

MFCC 实现的基本流程图如下图2.5, 按照预加重、分频加窗、离散傅里叶变换和三角空间滤波、离散余弦变换共五步。

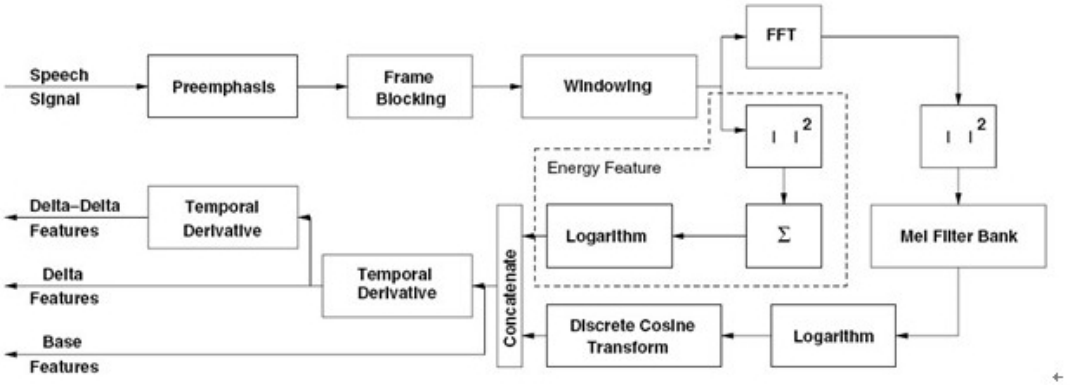


图 2.5　MFCC 实现流程图

1. **预加重**：由于音频信号在发声过程中会受到嘴唇等身体部位的干扰，需要通

过一种高通滤波器来强化高频部分。这样处理后的音频信号会显示的更加平坦，滤波公式为公式2.6

$$H(Z) = 1 - \mu z^{-1} \tag{2.6}$$

2. **分帧加窗**：由于手机的采样频率是 44.1KHz，即每秒采样 44.1K 个点，而一般采样的音频频率为 22kHz。又因为一个标准帧的长度为 25ms。所以每帧有 $22000 * 0.025 = 550$ 个采样点。分帧操作就是对每帧按照 10ms 的跨度进行拆分。假设声音信号为 $s(n)$, 完成分帧操作后为 $s_i(n)$;
3. **离散傅里叶变换**：对 $s_i(n)$ 做离散傅里叶变换得到 $S_i(k)$，对应的功率密度为 $P_i(k)$;

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \tag{2.7}$$

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \tag{2.8}$$

对分帧后的数据还需要按帧乘以汉明窗。这种方法可以显著提高帧左右段的连续性，实现公式如下

$$S_i(k) = S_i(k) \times W_i(k) \tag{2.9}$$

$$W_i(k, a) = (1 - a) - a \times \cos\frac{2\pi n}{N - 1} \tag{2.10}$$

这里的 a 一般取为 0.5

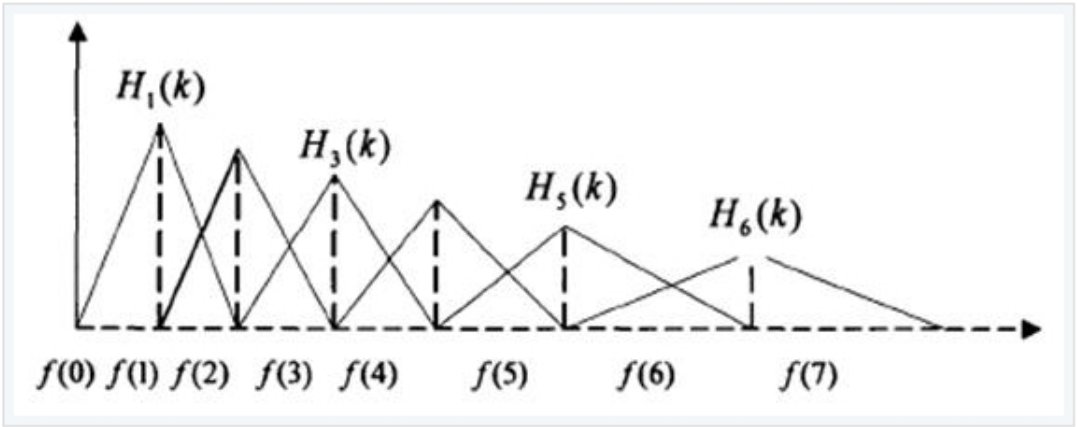


图 2.6　三角滤波器组的设置

4. **三角空间滤波**：通常使用 26 个左右的三角滤波器，对 $P_i(k)$ 进行滤波, 根据公式2.5, 可以求出 MEL 最大为 $f_{max}[mel] = 2146.1mel$, 由于三角空间滤波的

设置往往是等间隔的处理，所以其频率的间隔为公式2.11

$$\delta = \frac{f_{max}}{\kappa + 1} = 93.3mel \tag{2.11}$$

同时可对相邻三角空间滤波的上限频率进行设置

$$c(l) = h(l-1) = o(l+1) \tag{2.12}$$

具体如图2.6所示

5. **离散余弦轩变换**：在求离散余弦变换前需要计算每个滤波器组输出的对数能量

$$s(m) = ln(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)) \tag{2.13}$$

其中 $0 \leq m \leq M$, 之后利用 DCT 求得 MFCC 系数：

$$C(k) = \sum_{m=0}^{n-1} s(m) \cos \frac{\pi k(m-0.5)}{M} \tag{2.14}$$

由此对图2.2进行 mel 图谱的构建，其结果如图2.7所示

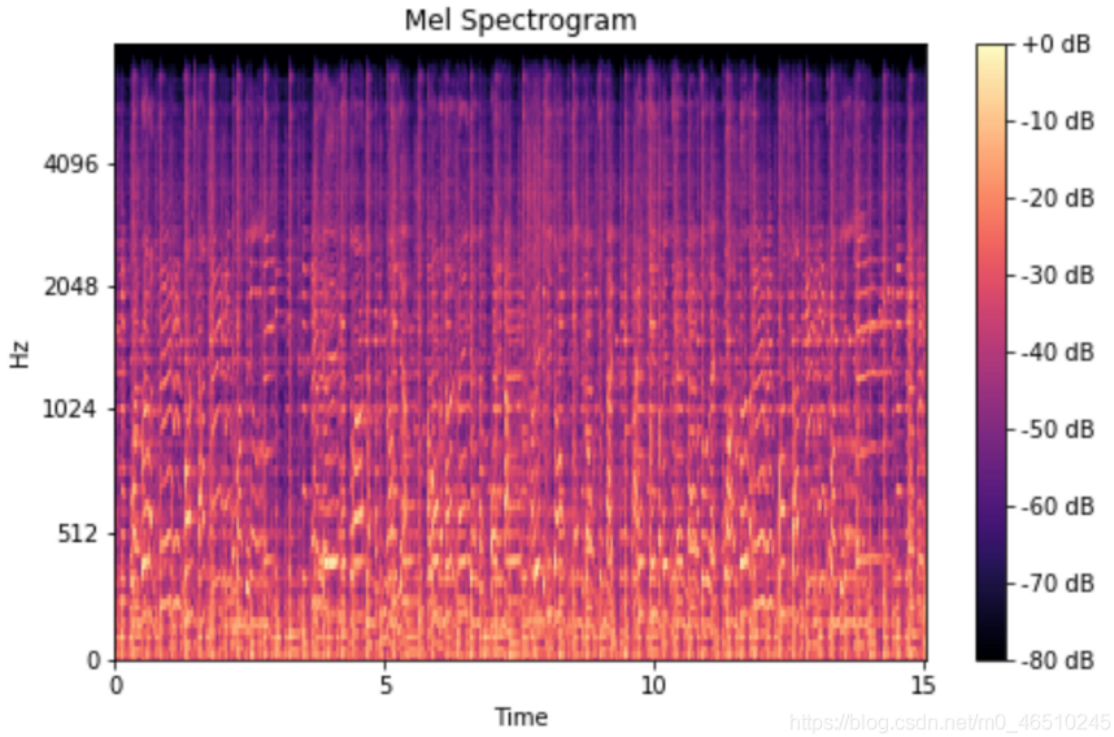


图 2.7　Mel 图

2.2 音频数据增强

2.2.1 概述

音频数据增强最早来自于图像神经网络中的图像增强方法，两者的目的都是为了拓展训练集并鼓励系统对增强过程中的转换保持不变。作为一种补充措施，对系统在转换后的输入上的预测进行学习可以提高针对系统未学习到的样本的鲁棒性。

在音频数据增强领域，现今的研究成果已经硕果累累了。2013 年，Jaitly 和 Hinton[11] 率先使用了保留标签的音频转换来进行语音识别。他们发现，在训练和测试时间进行 mel 滤波之前，频谱图的音调移位会使电话错误率从 21.6%降低到 20.5%，并报告称按时间或频率维度缩放 mel 频谱，或者根据扰动的 LPC 系数构造示例都无济于事。同时，Kanda 等人的研究成果[12] 表明，将音调移位与时间拉伸和随机频率失真相结合可将字错误减少 10%，而音调移位被证明是最有益的，并且三种失真方法的效果几乎呈线性相加。Xiaodong Cui 等人[13] 实现了将音高转换与在特征空间中将语音转换为其他说话者语音的方法结合起来的方法。

在这些研究的基础上，本研究可以确定音频数据增强在提高模型系统性能，避免过度拟合从而提高其通用性。

2.2.2 常用的音频数据增强方法介绍及展示

现今常用的数据增强方法有：加入高斯白噪音、音调转换、时间拉伸、时间平移以及同类音频叠加等。因此本文利用 github 上的音频数据增强库：audiomentations，展示一些常用的音频数据增强方法，并分析它们的影响。

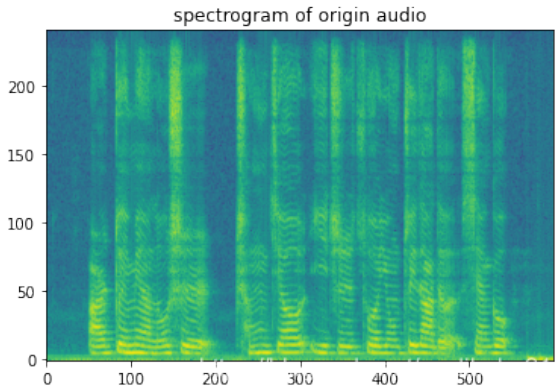
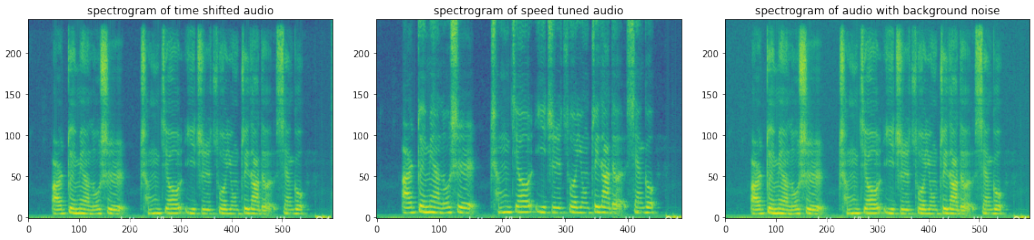


图 2.8　原始信号的频谱图

按照第一部分介绍的音频特征建模技术，音频数据增强的方法可以用于处理

展示频谱图和梅尔频谱图。因为梅尔频谱图需要构建复杂的三角滤波器组，所以本文用频谱图来分析展示数据增强方法，以及这种处理方法的影响。原始信号的频谱图展示如图2.8：

分别对原始信号做时间平移、时间拉伸和添加白噪音操作其频谱图变化如图：



(a) 时间平移　　(b) 时间拉伸　　(c) 添加白噪音

图 2.9　音频数据增强后的频谱图

在 audiomentations 库中，还有剪切样本、频段屏蔽、质量压缩等音频增强技术的实现。而这些音频增强的方法在最终的模型的影响，将在后面的实验设计过程中进行分析。

2.3　迁移学习

迁移学习是一种微调预训练模型，并快速标记适合的数据的深度学习方法。它与从零开始的深度学习训练相比，微调模型需要更少的标记数据，可以更快的实现训练。许多迁移学习任务的方法主要有：(i) 从一个预先训练的模型开始，(ii) 移除特定于任务的顶层，(iii) 作为特征提取器对目标任务的底层进行微调。通过这种方法，迁移学习系统可以实现特定任务的特征提取，而本文就采取利用 Youtube 数据训练的类神经网络 VGGish 实现基于咳嗽音频的特征提取。

2.3.1　研究现状

Donahue 等人在 2015 年发表的论文[14] 中确认经过预处理的 AlexNet 提取的特征可以转移到各种任务中，并且比手工制作的特征工作得更好。Yosinski 等人的研究[15] 表明，微调预训练网络比固定的预训练表征提供更好的性能。即使目标数据集与预先训练的数据集非常不同，微调没有带来性能提高，但可以加快收敛速度。

最近关于迁移学习的研究主要集中在如何更好地利用预训练模型的归纳偏差，

即如何利用预训练模型对微调进行规范化。在这些关于迁移学习的研究中都反映了迁移学习作为一种广泛的特征提取方法，有着很好的优势。

2.3.2 一般的迁移学习方法

一般来说，将某一领域的学习模型应用到相关领域是最常见的迁移学习过程。迁移学习方法主要有两种，分别是基于模型的迁移学习以及基于特征的迁移学习。

1. **基于模型的迁移学习**：基于模型的迁移学习主要是基于模型的再学习，经典的模型迁移学习方法就是 TrAdaBoost 算法，它主要通过增加误分类的目标函数，重新训练数据的权重，同时减少误分类。使得模型更加的适应新方向的数据集。
2. **基于特征的迁移学习**：本文主要就是使用基于特征的迁移学习，关注的是哮喘音频领域和人声特征领域的共同特征，然后利用这些特征进行特征映射。基于特征映射的迁移学习算法主要研究如何将源域和目标域的数据从原始特征空间映射到新的特征空间。这样，在这个空间中，源域的数据分布与目标域的数据分布是一致的，这样就可以更好地利用源域已有的标记数据样本在新的空间中进行分类训练，最终对目标域的数据进行分类测试。

2.4 卷积神经网络 VGG

2.4.1 概述

在传统的计算机机器学习中，特征提取往往是最重要的一环。通常机器学习程序员会对原始数据进行各种变形、处理，并人工的选取有代表性的维度。然后利用传统的机器学习分类网络，例如：支持向量机、决策树等进行分类。如果用于分类的特征不具有特异性或者特征点过少，分类器都无法进行有效的分类；而特征点提取过多，又会是分类系统对学习集过拟合，在新样本上的分类能力很差。

由于计算机性能的提高，GPU 和大型分布式集群的发展。近年来，卷积神经网络的研究越来越火爆，而利用神经网络系统就仅仅只需要关注网络层的功能和具体的参数限制，而不必太拘泥于数据样本的处理提取，利用网络自动地将自身需要的特征维度提取出来。尤其是 ImageNet 大规模视觉识别挑战赛（ILSVRC）[16] 已经用作多代大规模图像分类系统的平台，从而导致了深度视觉识别体系结构的许多进步。

本文将使用 Google 团队在 YouTube 数据集上预训练好的 VGGish 对咳嗽音频

进行特征自动提取，因此在本节有必要对卷积神经网络 VGG 的基本层和算法原理都需要进行简单的分析介绍。

2.4.2 VGG 的定义和基本层

Visual Geometry Group（VGG）是一类包含卷积计算并且有深度结构的前馈神经网络（Feedforward Netural Networks）[17]，主要应用于人脸识别、图像分类等方面。下面对卷积神经网络常有的四个基本层及其主要作用。

1. **卷积层**：该层主要用于提取输入数据的局部特征。对卷积核（特征提取器）和定义大小的数据执行卷积运算，并且输出层的结果值。一般来说，常用的卷积核有两种：一维卷积核和二维卷积核 1。卷积核的大小也决定了局部特征提取的大小。核太小会增加特征维数，核太大会使卷积提取特征的“粒度”不够精细。因此，选择合适的卷积核是非常重要的。
2. **池化层**：这一层主要是压缩特征维度和减少参数，而不会丢失太多的数据信息。在具体操作中，一般选择合适大小的窗口，并将所有特征值压缩成代表值。一般来说，有两种池方法：最大池和平均池。
3. **全连接层**：全连接层通常用作卷积层和输出层之间的连接，用于将多维特征值转换为所需的输出值。
4. **dropout 层**：一般添加到全连接层，减少中间特征的数量，防止模型拟合过多，提高模型的泛化能力。

基于以上四种基本层次从高维浅层特征编码到深层 ConvNet 编码，计算机视觉分类的研究在网络层数上越来越多。但是随着网络深度的增加，会出现更多由降维引发的问题，参数的过多也往往影响着机器学习的效果。

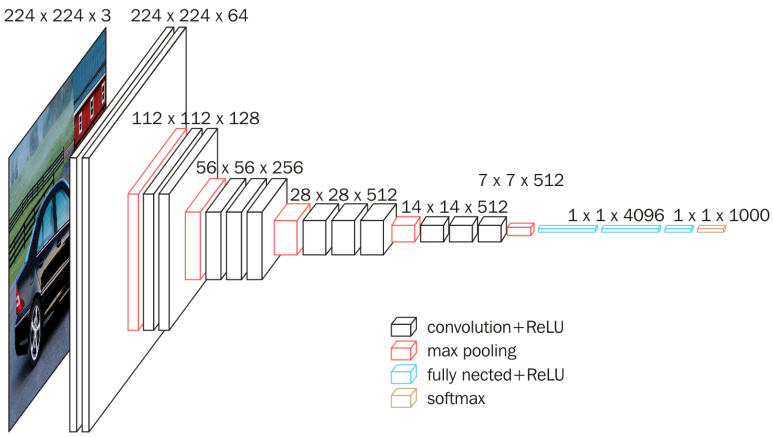


图 2.10　VGG16 结构示意图

由此，在 2014 年 ImageNet 大规模视觉识别挑战赛上牛津大学基于 AlexNet 网络提出 VGG 架构。相对于 AlexNet 网络，VGG 减少了其卷积核的大小。它利用 3 个 3×3 的卷积核来替代 7×7 的卷积核，使用了 2 个 3×3 的卷积核替代了 5×5 的卷积核[18]。这种小型的卷积核而不是大型卷积核可以保持更好的图像的性质,VGG16（图2.10）就是一种最经典的的 VGG 网络。

2.4.3 VGG16 网络结构及特点

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 2.11　VGG 常见结构配置

根据卷积神经网络中卷积核和卷积层数目的不同，如图2.11所示 VGG 共有六种常见的结构，其中后面两种结构的使用场景最多也最经典。D 结构就是本节主要介绍的 VGG16

通过对 VGG16 的结构进行具体分析，其组织结构一共包含以下层：

- 13 层卷积层，在表中以 conv3-YYY 标号
- 3 层全连接层, 在表中以 FC-YYY 标号
- 5 层池化层, 在表中以 maxpool 标号

其中,权重层为卷积层和全连接层,总共有 16 层,因此此种结构被称为 VGG16。

VGG 网络的最大特点就是在于简单，因为把所有的大卷积核都缩小为 3×3 的尺寸，并且按照若干层卷积层加上一层池化层的折叠方式，网络结构容易拟合。但是同时 VGG16 的权重参数有 1 亿多个，训练和调参难度过大，所以本文直接调

用利用 YouTube 音频数据训练的网络 VGGlish 提取参数。

3　系统设计

3.1　系统总体设计

本章将介绍基于音频特征建模的哮喘检测系统的实现细节。整个系统的设计架构如图3.1所示：在数据收集阶段，系统收集用户的咳嗽信号后，利用数据增强音频数据，并通过人工音频特征提取和 VGGish 特征提取。最后将提取到的特征输入 SVM 分类网络进行分类训练，根据分类指标评估模型的好坏。

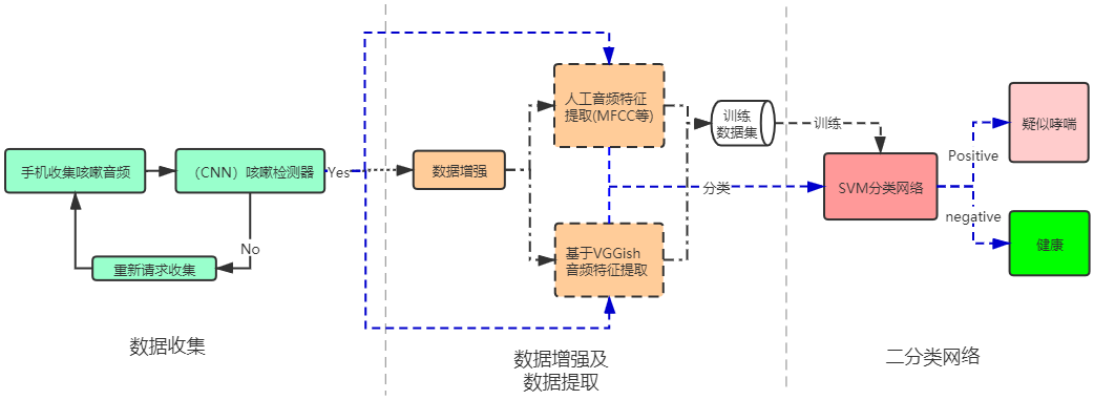


图 3.1　哮喘检测系统整体架构

本章主要介绍哮喘检测系统的四个核心部分：(1) 利用收集麦克风收集咳嗽音频，并利用 CNN 神经网络对咳嗽音频的收集结果进行评估；(2) 在对咳嗽音频进行音频数据增强后，利用人工处理和 VGGish 网络进行联合特征提取；(3)SVM 分类网络，介绍用于本系统的 SVM 分类网络架构，简述原因以及分类效果。

3.2　原始数据收集及检测

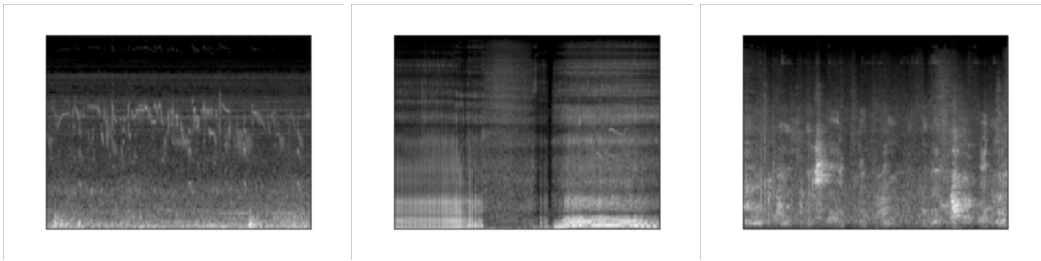
系统主要基于 Android Studio 平台搭建了一款收集咳嗽音的手机应用，通过调用手机的麦克风以 44.1kHz 的采样率对咳嗽音频进行收集，并转化为.wav 文件存储到手机本地内存中。但是系统并不能确定原始收集数据是否为咳嗽音频，每一次调用的录音功能都不能保证收集到的数据一定是可用的咳嗽样本。为了使样本满足可用性这一条件，系统需要设计一个简单的 CNN 网络用于咳嗽样本的检查。

3.2.1 咳嗽样本预处理

根据理论基础中的 MFCC 理论，系统利用生成咳嗽样本的 Mel 图用于图像识别分类，处理过程如下：

1. 将咳嗽音频重定位至 22kHz。
2. 使用 128 个 Mel 成分的三角滤波器对咳嗽音频进行滤波并确定其 Mel 图。
3. 调整已经生成的 Mel 图大小并转化为灰度图，以统一的大小对图像进行缩放并减小图像尺寸，生成 $320 \times 240\times$ 大小的灰度图。
4. 将生成的图像输入到基于系统的卷积神经网络（CNN）的分类器中，以确定记录的输入声音是否咳嗽

经过系统预处理后的音频图像是一个 $320 \times 240 \times 1$ 的灰度图，系统用于训练的图展示如图3.2：其中图3.2a为咳嗽音频图像，图3.2b和图3.2c为其他音源的音频图像。



(a) 音频灰度图 1　　(b) 音频灰度图 2　　(c) 音频灰度图 3

图 3.2　咳嗽检测训练集样本预处理

3.2.2 CNN 网络结构

CNN 网络的结构设计如图3.3, 其中包含 4 个卷积层、3 个最大池化层以及 1 个激活层。最后神经网络生成的 2 个神经元和 softmax 激活层用于进行咳嗽与非咳嗽的分类。

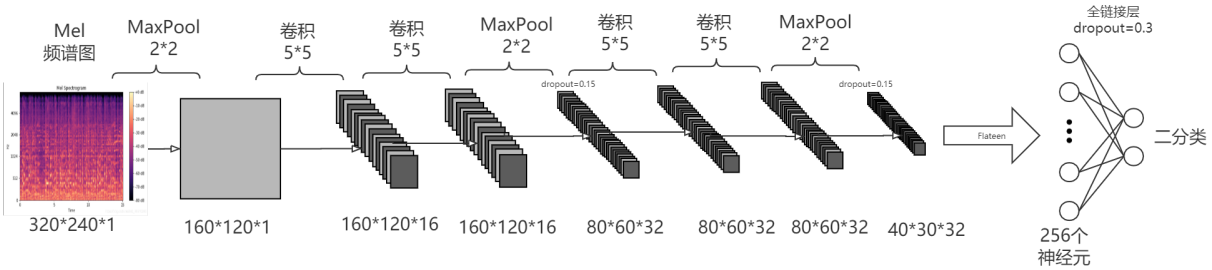


图 3.3　咳嗽检测 CNN 网络

CNN 网络的设计思路如下：

1. 第一部分是最大池化层。由于输入的 Mel 光谱图图像尺寸较大，因此在进行下一步之前，首先将其穿过 2×2 最大池化层以减小整体图像的尺寸，方便进行卷积等特征提取处理。
2. 第二部分是两个图层块。每个图层包括两个卷积层, 并将卷积核大小设置为为 5×5，用于提取输入数据的局部特征。其中，第一部分的图层中的卷积层使用了 16 个过滤器，而第二部分中的两个卷积层使用 32 个过滤器。系统采用较大的 5×5 的卷积核是为了减少造成局部特征的过拟合，提高了整体网络的泛化性能与指标。
3. 第三部分是每个图层块后面的一个 2×2 的最大池化层以及一个系数为 0.15 的 Dropout 层。最大池化层是为了进一步减小图像尺寸，而 Dropout 层是为了随机让网络某些隐含层节点的权重不对合并进行产生影响，提高整体神经网络的可靠性。
4. 第四部分是一个完全连接层。系统将从这 4 个卷积层中学习到的复杂特征进行展平处理，传递到具有 256 个神经元的完全连接层，最后利用系数为 0.3 的 Dropout 层继续防止过拟合情况的发生。
5. 最后一部分是 2 个神经源和 softmax 激活功能的输出层。通过这层结构系统就可以完成对给定输入的咳嗽与非咳嗽音频进行分类。

同时对于该神经网络系统还有一些实现的细节。在该模型中，系统使用 ReLU 函数作为所有卷积层的激活函数，并利用 Adam 优化器优化神经的网络结构[19] 以及使用交叉熵损失函数作为卷积网络的损失函数。

3.3 特征提取

由于本文使用的数据量较小，系统将采用基于特征的机器学习来提取音频信号的特征向量。系统实现了传统的手工设计特征和迁移学习两种不同的特征提取方法，并使用主成分分析法对系统所有提取到的特征进行降维处理。

3.3.1 手工设计特征

对于咳嗽音频接下来系统将指明本文挑选的特征值以及挑选的原因

1. **短时平均过零率（Zero Crossing Rate）**:ZCR 是指一定时间内音频波形通过直线 y=0 的次数。这个给特征值在一定程度上反映了频率的大小。ZCR 低的

音频一般比较浑浊，而 ZCR 高的音频则比较清澈，因此短时平均过零率用于初步分析清晰、浑浊的音频。通过因为短时平均过零率容易受到低频噪音的干扰，为了提高鲁棒性，系统会在处理中添加阈值，即波形穿过阈值的次数被定义为短时平均过零率。计算算法如下：

Algorithm 1 ZCR 计算

Require:

采样率：fs, 一段时间的音频信号：f(t), 帧长：wlen, 帧移：inc, 阈值：α

Ensure:

消除直流分量 f(t) = f(t)-mean(f(t))
初始化 wlen = 200, inc = 80, count=0, α=0.05*max(f(t))
分帧 F, 获取帧数 fn
for i = 1:fn **then**
z=F[i]
for i = 1:(wlen-1) **then**
if z(j)*z(j+1)<α **then**
    count++;
end
end
end
return count

2. **短时音频能量**：短时音频能量是音频信号相应的帧长归一化的振幅值的平方和，依然是用来分辨清澈和浑浊音频的的指标，在时刻 T，其计算公式如下：

$$E_n = \sum_{m=t-(N-1)}^{t} F[m]^2 \tag{3.1}$$

3. **能量熵**：系统将子帧归一化后取其能量的熵，用来度量音频中突变音的发生。
4. **共振峰频率**: 共振峰是由人声道的共振引起的频谱整形。
5. **峰度**: 峰度是对实值随机变量的概率分布的右部分的度量。
6. **RMS 能量**：RMS 能量是信号功率的短时傅里叶变换幅度的均方根，用于表征信号中的能量大小。
7. **频谱质心**：功率谱图每帧功率的平均值（即质心）。
8. **截止频率**：功率谱图的中心频率，以便此帧中至少 85% 的频谱能量包含在该值及以下。
9. **MFCC**:MFCC 反映了频谱的轮廓，系统基于非线性 Mel 尺度上对数功率谱

的线性余弦变换，从短期功率谱获得 Mel 频率倒谱系数。系统使用前 13 个分量的值。

10. **Δ-MFCC**:MFCC 的时间微分
11. **Δ^2-MFCC**:MFCC 增量的微分（加速度系数）

对于产生的具有时间性质的特征（如：均方根能量、频谱质心、截至频率和 MFCC 的所有变体），系统提取了一些统计特征，以便捕捉超出平均值的分布。包括：平均值，中位数，均方根，最大值，最小值，1/4 位数和 3/4 位数，四分位的间距，标准差，偏度，峰度。总共有 477 个手工设计特性，包括 3 个段级特性、4 个由其统计数据表示的帧级特性和 3 个 MFCC 的特征，每个组件由其统计数据表示（3+4）$\times$ 11 + 3 $\times$ 13$\times$ 11 = 516。

3.3.2 迁移学习特征

除了手工设计的特征外，系统还使用 VGGish 来自动提取音频特征 [15]。VGGish 是一个卷积神经网络，主要用于基于原始音频输入的音频分类。VGGish 是使用大型 YouTube 数据集进行了训练的模型，并公开发布了学习到的模型参数。因此本系统将其用作特征提取器，将原始音频波形转换为嵌入（特征），然后将其传递以训练 SVM 分类器。具体训练方法如下：

1. 将音频重采样为 16kHz 单声道,
2. 使用 25ms 的帧长、10ms 的帧移，以及 Hann 窗口对咳嗽音频进行分帧，切割为 0.96s 的非重叠子帧
3. 对每一帧做傅里叶变换，然后利用信号幅值计算声谱图
4. 通过将声谱映射到 64 阶 mel 滤波器组中计算 mel 声谱, 并对声谱图取对数能量
5. 模型每 0.96 秒返回 128 维特征向量，将整个线段的均值和标准差作为最终特征，尺寸为 256（128$\times$2）

需要注意的是由于 VGGish 仅基于频谱图输入，因此时域中的一些重要特征可能会在特征空间中遗漏。

3.4 支持向量机

3.4.1 概述

支持向量机（support vectormachines）是机器学习中最经典的二分类模型。这种分类器提供了一种有监督的学习模型，该模型在训练后得到一个最大分离两类的超平面决策边界。

一般来说，为了处理非线性可分的数据，SVM 可以加入适当的核，将原始特征空间转换为高维空间，在高维空间中，转换后的特征成为线性可分的 (见图 4 的解释说明)。形式上，选择决策超平面 $w^T\varphi(x)+b=0$(其中 φ(x) 表示变换特征空间中的一个点向量，φ 为核函数，w 为权向量，b 为偏差) 来最大化整体分离。相当于最大化公式3.2

$$\sum_{i=1}^{n}\alpha_i - 0.5\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{3.2}$$

其中 $a_i \geq 0, i=1,2,...n$, 且 $\sum_{i=1}^{n}\alpha_i y_i = 0$。对于每一个数据的索引 i，$x_i, y_i$ 分别表示特征向量和对应类的标签（+1 代表正样本，-1 代表负样本），并且 K 表示核函数，n 表示数据集的大小，α_i 代表拉格朗日乘数。

本理论介绍两类支持向量机的分类性能，其中一种向量机的核函数为线性核（即：$K(x_i,x_j)=x_i^T x_j$），另一种的核函数是径向基函数（即 $K(x_i,x_j)=exp\frac{|x_i-x_j|^2}{2\sigma^2}$，这是一种泛在非线性核。

3.4.2 性能表现

为了评估分类性能，SVM 参数，即权重向量 w 和偏差 b，在数据子集 (训练子集) 上进行优化，并针对互补 (测试) 子集进行验证。准确地说，数据集被划分为训练和测试的子集，这样:(1) 他们大小的比率，称为训练-测试比率，这是一个预先分配好的参数；(2) 这些子集中健康以及患病的比例也大致于训练-测试比率相同。一般情况下，分类器的性能取决于主观选择的分割条件。为了避免性能分析中这种主观性，通常使用蒙特卡罗交叉验证 (MCCV)[20]，如图3.4所示, 本文的所有数据集被随机划分大量 (5000) 次 (迭代)，并分割比和前面提到的训练-测试比例保持不变。在每一次迭代，SVM 参数在训练子集上进行优化，并记录平均训练、测试精度和相应的标准差。

值得注意的是，训练精度表明了当前模型对可见数据的分类性能，而测试精度表明，对于不可见数据，平均测试精度高的分类器具有实际意义。此外，低标准差

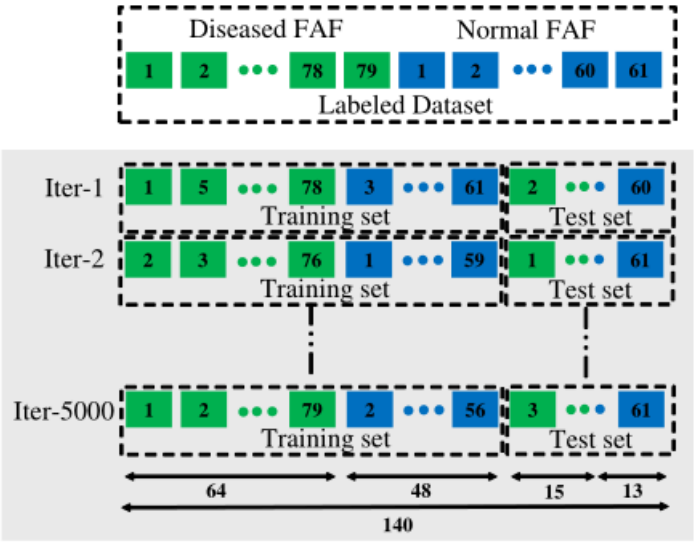


图 3.4　MCCV 方法流程

(该情况表明在随机分区上的低可变性，从而表明系统的健壮性) 是可取的。并且通过计算了超过 5000 次迭代的平均混淆矩阵，系统提供了健康类和疾病类的每个类条件检测概率。最终实验环节实验组训练和观察了 SVM 分类器不同的训练-测试比率对 SVM 性能的影响。

4 实验设计及其结果

4.1 实验准备

为了成功完成哮喘检测实验，系统需要准备的工具有：Android 应用程序（用来收集咳嗽音频，并存储在本地内存中）以及装有 tensorflow 和 matlab 环境的笔记本电脑（用来提取音频特征并进行对音频进行分类）

在实验数据采集阶段，本人联系了武汉市中南医院呼吸科找到了 4 名患有哮喘的病人分别采集了 10min 左右的咳嗽信号，并利用咳嗽检测网络筛选了 300 组有效的单周期咳嗽信号，作为本系统的正样本数据。紧接着本人联系了 5 名身体健康的志愿者共收集了 500 组负咳嗽样本数据。

4.2 咳嗽检测网络测试

对应已经收集到的咳嗽信号，系统要对咳嗽事件进行正确的检测，去除样本中噪音大或者非咳嗽的样本，因此系统需要训练一个鲁棒性良好的咳嗽检测网络。

本网络的数据是使用 Kvapilova L 等人[21] 的采集咳嗽样本作为正样本进行训练，将咳嗽样本与一些背景噪声混合以使其更加真实，从而达到扩大数据集的目的，背景噪音来自于 musan 数据集[22], 其中噪音比为 0.15。最终系统利用在数据采集阶段收集到的咳嗽样本（包括正常人咳嗽样本和病人咳嗽样本）对真实的样本进行分类检测。来观察训练出的网络对真实样本分类的有效性。

4.2.1 参数重要性分析

系统利用初步网络训练对参数的重要性和相关度进行分类。

该表显示了不同的超参数对 “真实样本准确率” 的影响。第一列显示了参数的重要性，第二列显示了与参数更改量之间的相关性，以及参数增加是否改善或恶化了参数。通过准确性得分，系统可以看到 β_1、β_2 以及 α 对系统的影响很小，因为没有任何变化带来了性能提升。但同时几个卷积层以及损失函数对本网络的数据影响很大，但尚未经过足够的测试，应对此进行调查。

参数	重要性	相关度
drop1	**0.041**	**0.243**
conv3	**0.063**	**0.133**
conv2	0.071	0.124
drop2	**0.038**	**0.091**
conv4	**0.047**	**0.089**
batch_size	0.041	0.041
lr	0.048	0.028
conv1	0.050	-0.021
pool1	0.432	-0.113
beta2	0.013	-0.121
alpha	0.050	-0.257
beta1	0.106	-0.274

表 4.1　模型参数重要性分析

4.2.2　模型参数调整

本节为对于咳嗽检测网络进行网络模型训练的参数调整确认。通过调整系统的参数，实验验证了不同的系统参数对于模型的影响。

由上节的分析系统可以看出参数中卷积层的大小以及损失函数对模型的效果影响最大。所以需要选取适宜的网络的损失函数和子图数目。通过初步筛选，系统将损失函数确定为：交叉熵损失函数和最大似然损失函数两种。假定训练时所有条件均保持不变，按照第三章的网络结构进行设计，即在卷积核大小为 5×5，池化核大小为 2×2，训练集为 1800，batch 数目为 128，训练次数为 50，优化器采用 Adam 的条件下进行训练对比，其结果如下：

损失函数/子图数	验证集准确率	训练集准确率	真实样本准确率
交叉熵/（16，16，32，32）	**0.9756**	**0.9673**	**0.8941**
交叉熵/16，16，16，32	0.9732	0.9682	0.8888
最大似然/16，16，32，32	0.9672	0.969	0.8574
交叉熵/32，32，32，32	0.9731	0.9663	0.862
最大似然/16，16，16，32	0.9142	0.9008	0.708
最大似然/32，32，32，32	0.9678	0.9669	0.8914

表 4.2　调整损失函数和子图数对模型的影响

通过表格系统可以看到当系统使用交叉熵函数作为系统的损失函数且子图数选取为 16，16，32，32 时样本分类准确性和真实数据分类准确率最高，到达了 89.4% 基本适用于进行咳嗽的检测。

4.2.3 模型决策修正

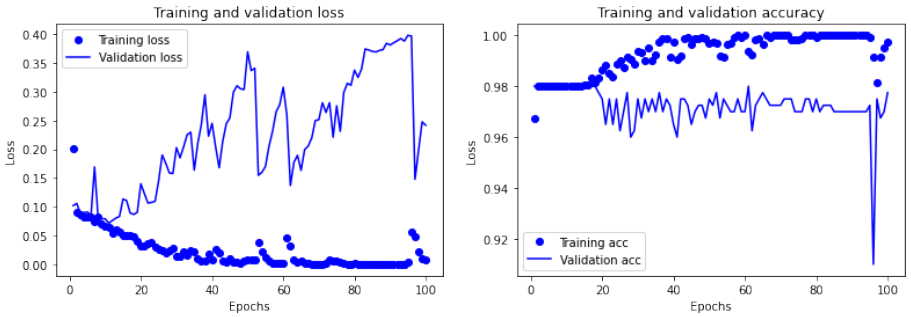
在模型参数调整阶段，系统已经将真实样本的决策准确率提高到 89%。系统可以知道单周期的咳嗽音频信号可能难以特别准确分类（到达 99%）以上，所以说系统修改分类判断，对一段时间内多次的连续的咳嗽信号进行联合分类判断，用于修正模型的决策过程, 具体理论以及决策过程如下：

1. 存储连续 N 个连续的咳嗽音频信号，进行单周期的分类决策。
2. 如果 N 个信号中，有 T 个信号及其以上的信号判断一致时，系统采取该判定值为本次连续分类决策的最终结果。
3. 否则回到第一步，重新存储 N 个连续信号，进行联合分类决策

在上述的理论情况下，系统使用公式4.1估算当 N=3,T=2 时联合分类决策的错误率：

$$ErrorRate = C_4^3 \times (0.11)^3 \times 0.89 + 0.11^4 \approx 0.488477\% \tag{4.1}$$

以上理论表明，即使单周期咳嗽音信号的识别率在 89% 左右，但是通过四次连续的联合分类决策，可以将识别率提升到 99.5% 左右，系统按照此方法对网络进行训练以及预测，其网络结果如4.1。最终的分类成功率在 99% 左右。



(a) 训练阶段损失函数变化　　(b) 训练阶段准确率变化

图 4.1　CNN 训练过程展示

4.3 支持向量机网络测试

特征提取系统结构中论述所示，系统通过人工设计特征以及迁移学习获取了 700 多维度的特征向量并使用 PCA 降维最终形成长度为 118 的特征向量。如前所述，系统考虑了两种不同的支持向量机分类器，线性支持向量机和带 RBF 核的支持向量机 (RBF-SVM) 进行性能比较。在每一种情况下，训练和测试的准确性值被记录在大量 (5000) 的随机分区中，对于在 10:90 和 90:10 之间选择的各种训练测试

的划分比，其平均值 (标准差在括号中) 列于表中。

表 4.3　样本相同，不同训练测试集划分比，不同 SVM 的分类结果

训练-测试集合比率	线性 SVM		RBF-SVM	
	训练集准确率	测试集准确率	训练集准确率	测试集准确率
10:90	91.27	75.03	99.38	78.97
20：80	91.51	82.26	98.81	84.32
30:70	92.04	85.56	97.88	86.66
40:60	92.26	87.31	98.86	88.11
50:50	92.24	88.40	98.86	88.89
60:40	92.27	89.02	98.46	89.58
70:30	92.31	89.53	98.62	90.10
80:20	92.26	89.88	98.65	90.55
90:10	92.25	89.60	**98.86**	**90.83**

显然，对于线性支持向量机，平均训练和测试准确率水平随着训练测试比的增加而增加，但也有少数例外。正如预期，随着训练数据的可用性的增加，该分类器往往学习得更好。对于 RBF-SVM 模型中，系统所得到的平均测试精度，仍然很大程度上遵循上述增加的趋势，比使用线性支持向量机获得的每个分裂比略有改善。这表明了潜在问题中固有的中等程度的非线性。尽管在测试精度上只有轻微的提高，但平均训练精度的提高明显更高，这表明 RBF 核在建模方面训练数据的某些非线性方面，而这些方面不能很好地一般化划分。

实际上，考虑标准差值也可以得出类似的结论如下。在线性 SVM 和 BBF-SVM 两种模型中训练精度的标准差都随着划分比的增加而减小，并且后者的值明显更低，表明 RBF-SVM 建模效果更好。然而，从测试精度来看，当划分比值较低时，RBF-SVM 的标准差较低，当划分比值较高时，线性 SVM 的标准差较低。这一现象表明系统泛化性能较差。随着划分比的增加，两种情况下的标准差先减小后增大可以看出细微的差别。而测试的准确性是最可靠的 (即，具有最低的标准偏差)，在线性 SVM 的情况下 (也在 [26] 其他地方观察到) 一个均匀的划分训练测试比，在 RBF-SVM 的情况下，最高的可靠性是在划分比 40:60 观察到。

虽然对于健康类，线性支持向量机在准确性和可靠性方面略优于 RBF-SVM。请注意，给定疾病类别的 RBF-SVM 相对于线性 SVM 的优势与给定健康类别的线性 SVM 相对于 RBF-SVM 的优势是相似的。然而，由于给出患病类比给出健康类会产生更高的实际代价，因此应该选择 RBF-SVM 作为筛选工具。

综上考虑系统用 80：20 划分条件下，RBF-SVM 的支持向量机。此时 RBF-SVM

	预测\真实	哮喘	健康
线性 SVM	哮喘	89.47	7.75
	健康	10.53	92.25
RBF-SVM	哮喘	91.61	11.41
	健康	8.39	88.59

表 4.4　训练测试集划分比为 80:20 的预测混淆矩阵 (%)

对健康类的正确率为 **88.59%** 对于哮喘类的正确率为 **91.61%**，标准差分别为 **7.09%** 和 **8.82%**。

5 总结与展望

5.1 论文工作总结

在本文中系统研究了基于咳嗽音频的哮喘检测方式，由于系统的检测方式是基于 Android 平台搭建的，而不需要专业的检测机器，因此对比其他很多的哮喘诊断方法具有更好的实用性和普查性。系统再此研究的背景下研究了音频中咳嗽事件的检测、迁移学习实现特征提取以及支持向量机份分类的性能问题。

对于音频中咳嗽事件的检测问题，系统提出建立一个 CNN 网络进行咳嗽事件的检测。该网络是利用 Kvapilova 等人数据库的采集咳嗽样本加以环境噪音进行训练的。系统通过分析网络参数的重要性和相关性后，将网络性能的重点聚焦于特征子图的个数以及损失函数的选取。最终通过分析确定了损失函数为交叉熵损失函数，特征子图个数为 16，16，32，32，并经过联合分类决策，极大地提高了检测网络的可靠性。

对于迁移学习问题，系统使用 Youtube 数据集训练的类神经网络 VGGish 提取咳嗽音频数据的特征向量，但是因为 VGGish 仅基于频谱图输入进行特征提取，会遗漏时域中的一些重要特征，因此系统还采用了人工设计的方式提取了新的 500 维特征向量，保证了特征空间的全面性和准确性

最后对于支持向量机的分类性能问题，系统主要分析了线性支持向量机和带 RBF 核的支持向量机 (RBF-SVM) 的性能比较关系。由于系统提取到的特征向量维度较高且重点不突出，系统采取 PCA 方法对特征向量进行降维处理，并分析不同训练比对两类向量机性能的影响，最终系统选取了训练集-测试集 80:20 的情况并采用 RBF-SVM 进行特征分类，达到了 90% 以上的分类效率。

综上所述本研究主要实现一种基于咳嗽音频的哮喘检测方式，通过手机麦克风收集用户的咳嗽音频，并在经过音频数据增强后，利用联合特征提取和 VGG 模型特征提取，提取一个 733 维度的特征向量，将所有的特征点进行主成分分析法（PCA）降维处理后，使 RBF 支持向量机进行二分分类，分类准确率平均达到 90% 以上。

5.2 下一步工作

虽然本论文的最终分类准确率达到了 90% 以上，但是本研究的局限性也很大。首先是音频提取的负样本不够全面，没有办法很好的涵盖各种环境下咳嗽音的检测，在环境噪音较大的情况下决策失败率很高，需要多次采样。未来改进的方法主要有以下两种：1. 使用 VGGish 模型对咳嗽事件检测采取小样本学习；2. 继续加大负样本量，通过数据增强等方式，提高模型的鲁棒性

其次在哮喘检测的工作中，只是简单的对哮喘健康进行分类，而没有对其他异常咳嗽情况进行检测和分类。在测试样本中，有一名志愿者没有患有哮喘但是近期有咽炎症状，分类效果并不良好，因此在上面的分类实验中，已经去除该志愿者的实验数据。为了提高系统的冗余性，该研究的下一步工作就是去寻找其他几种常见异常咳嗽病人进行细分类。

参考文献

[1] LOUIS R, LAU L C, BRON A O, et al. The relationship between airways inflammation and asthma severity[J]. American journal of respiratory and critical care medicine, 2000, 161(1): 9-16.

[2] VAN SCHAYCK C, CHAVANNES N. Detection of asthma and chronic obstructive pulmonary disease in primary care[J]. European respiratory journal, 2003, 21(39 suppl): 16s-22s.

[3] PRAMONO R X A, BOWYER S, RODRIGUEZ-VILLEGAS E. Automatic adventitious respiratory sound analysis: A systematic review[J]. PloS one, 2017, 12(5): e0177926.

[4] XU X, YU J, CHEN Y, et al. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals[C]//Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. 2019: 54-66.

[5] Zhang G, Mei Z, Zhang Y, et al. A noninvasive blood glucose monitoring system based on smartphone ppg signal processing and machine learning[J]. IEEE Transactions on Industrial Informatics, 2020, 16(11): 7209-7218.

[6] BARRY D T, HILL T, IM D. Muscle fatigue measured with evoked muscle vibrations[J]. Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, 1992, 15(3): 303-309.

[7] BALES C, NABEEL M, JOHN C N, et al. Can machine learning be used to recognize and diagnose coughs?[C]//2020 International Conference on e-Health and Bioengineering (EHB). IEEE, 2020: 1-4.

[8] SCHLÜTER J, GRILL T. Exploring data augmentation for improved singing voice detection with neural networks.[C]//ISMIR. 2015: 121-126.

[9] MCFEE B, RAFFEL C, LIANG D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference: volume 8. Citeseer, 2015: 18-25.

[10] ZHENG F, ZHANG G, SONG Z. Comparison of different implementations of mfcc

[J]. Journal of Computer science and Technology, 2001, 16(6): 582-589.

[11] JAITLY N, HINTON E. Vocal tract length perturbation (vtlp) improves speech recognition[C]//2013.

[12] KANDA N, TAKEDA R, OBUCHI Y. Elastic spectral distortion for low resource speech recognition with deep neural networks[C]//2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013: 309-314.

[13] CUI X, GOEL V, KINGSBURY B. Data augmentation for deep neural network acoustic modeling[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(9): 1469-1477.

[14] EVERINGHAM M, ESLAMI S, GOOL L V, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111 (1): 98-136.

[15] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[J]. MIT Press, 2014.

[16] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

[17] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[19] KINGMAN D, BA J. Adam: A method for stochastic optimization. conference paper [C]//3rd International Conference for Learning Representations. 2015.

[20] LIANG X. Monte carlo cross validation[J]. Chemometrics and Intelligent Laboratory Systems, 2001.

[21] KVAPILOVA L, BOZA V, DUBEC P, et al. Continuous sound collection using smartphones and machine learning to measure cough[J]. Digital biomarkers, 2019, 3(3): 166-175.

[22] SNYDER D, CHEN G, POVEY D. MUSAN: A Music, Speech, and Noise Corpus [M]. 2015.

致谢

时光飞逝，四年的学习生涯即将画上圆满的句号。匆匆四年，我不仅学到了专业知识，还全面提高了自己的能力。人生之路是漫长的。我很珍惜在学校里帮助过我的每位老师和同学。

在完成毕业设计的过程中，首先要感谢吴渊师兄的帮助，耐心地指出了我毕业设计中存在的问题，并提出了许多有价值的建议，使本毕业设计原本完成。张健教授认真严谨的学术态度和敬业的工作精神为我今后的工作生涯树立了榜样。其次，我要感谢我最好的朋友官千云的帮助。在他们的帮助下，我成功地解决了问题，并按计划完成了我的毕业设计。

最后，我希望所有同学都能实现他们的梦想，以梦为马，不负韶华。