《检测技术及系统设计》

研讨报告

研讨题目：室外公共场所突发安全事件声学检测系统

院　　系：仪器科学与工程学院

专　　业：测控技术与仪器

小组成员：招梓枫　　学　　号：22017327

林　涵　　22017319

本科生课程考试成绩单

<table>
<tr><td>院　　系</td><td colspan="2">仪器科学与工程学院</td><td>专业</td><td colspan="3">测控技术与仪器</td></tr>
<tr><td>学生姓名</td><td colspan="2">招梓枫
林涵</td><td>学号</td><td colspan="3">22017327
22017319</td></tr>
<tr><td>课程名称</td><td colspan="6">检测技术与系统设计（Seminar）</td></tr>
<tr><td>授课时间</td><td colspan="2">2020 年 2 月 — 2020 年 6 月</td><td>周学时</td><td>32</td><td>学分</td><td>2</td></tr>
<tr><td>简
要
评
语</td><td colspan="6"></td></tr>
<tr><td>研讨题目</td><td colspan="6">室外公共场所突发安全事件声学检测系统</td></tr>
<tr><td>成绩</td><td colspan="6"></td></tr>
<tr><td>备注</td><td colspan="6"></td></tr>
</table>

任课教师签名：_____

日期：

摘要

随着信息化技术和物联网浪潮的快速发展，目前，基于摄像头等手段的公共场所安全监控已经十分普及和完善，大大提升了公共场所发生的突发事件的管控效率。然而，由于摄像头只能采集基于视觉的信息，视频监控不可避免地存在一些先天性缺漏。相对于纯视频手段，声学检测在枪击、爆炸、呼救、人群恐慌等具有语义的事件中，无疑具有更好的检测识别能力，这使得基于声学的公共安全事件检测在反恐、维稳、社会治安等多个领域具有广泛的使用价值和应用前景。本课题重点针对枪击与爆炸两类突发公共安全事件，对相应的声学检测方法进行了研究。

本课题结合室外公共场景下的枪击与爆炸事件声学传播情况、系统设计成本以及检测算法的算力开销等因素，对比了各类传声器在该场景下的选型优劣性，并基于 VM2020 麦克风、AD7865-1 模数转换模块、SRAM 拓展 IS61LV25616AL、FLASH 拓展 SST39VF800、TMS320F2812 数字信号处理器等设计了信号检测的硬件平台。

本课题针对室外公共场所的道路街道场景，设计了基于 Butterworth 低通滤波器的滤波去噪算法，和基于短时能量分析与持续时间分析的端点检测算法，实现了大功率前景信号与低功率背景信号的分离。为了对所分离的信号片段用于进一步检测分类，本课题采用 Mel 倒频谱系数（MFCC）作为特征工程方案，并应用混合高斯模型（GMM）与极大似然估计对疑似信号片段进行检测与分类。最后，本课题还结合 TUT Acoustic 与 Freesound 两个声学数据库的 60 个枪击、爆炸以及道路街道场景样本，通过 MATLAB 仿真进行了小规模算法验证。

除此之外，本课题还对有关的技术问题与系统设计问题进行了回答和解释，并在最后对进一步的研究方案及其设计可行性进行了简要的讨论。

关键词：枪声检测；公共安全；声学检测硬件平台；短时能量分析；MFCC；混合高斯模型；

目 录

第 1 章 绪论.....5
1.1 选题背景.....5
1.2 研究现状.....5
1.3 课题工作.....6
第 2 章 传声器选型.....8
2.1 选型要求.....8
2.2 指标概述.....8
2.3 需求转化.....10
2.4 种类选型.....10
第 3 章 配套硬件系统设计.....13
2.1 DSP 选型.....13
2.2 ADC 选型.....14
2.3 存储拓展与电源设计.....14
第 4 章 信号处理前端.....16
4.1 滤波降噪.....16
4.2 端点检测.....17
第 5 章 特征工程与分类器设计.....22
5.1 特征工程.....22
5.2 分类器.....24
第 6 章 展望与答疑.....28
6.1 进一步研究方向.....28
6.2 问题解答.....29
参考文献.....34

第一章．绪论

本章主要讨论选题背景、研究现状和课题主要工作，意在探讨选择声学手段作为检测方法的原因，以及具体的检测系统设计和研究路线。

1.1 选题背景

公共安全问题是社会安全稳定所聚焦的话题之一。近年来，检测技术与监控自动化正深刻地改变着人们的生活。尤其在安防领域，闭路电视 CCTV（Closed Circuit Television）、视频流分析、智能监控等新技术得到了广泛应用，大大提高了安防监控的管理效率。然而值得注意的是，基于视频流的监控手段不可避免地也具有一定的先天性缺漏，例如存在视野盲区、易受光照影响等问题，对于事件检测，还可能存在语义不明的问题，监控手段不够全面。纯视频手段在枪击、爆炸、暴恐袭击、人群恐慌等具有较强语义性的突发公共安全事件中，往往不如声学检测分析手段敏感和有效。声学事件检测主要是使用一些声学处理方法，刻画现场音频流的声学特征，再结合适当的分类器进行检测分类，从而实现对音频流中出现的声学事件进行检测分析。基于声学的公共安全事件检测在反恐、维稳、社会治安等多个领域具有广泛的使用价值和应用前景。本课题重点针对枪击与爆炸两类突发公共安全事件，对相应的声学检测方法进行了研究。

1.2 研究现状

对于枪声的研究聚焦在对于膛口波和弹道波的研究分析上。吴松林等[1]深基于弹丸的空气动力学模型，深入分析了弹道激波的成因和理论波形；蒋灏等[2]分析了小口径武器发射的膛口波和弹道激波，并设计了基于膛口激波的 DOA 模型对弹丸弹道轨迹进行估计；卢慧洋[3]分析了弹道波和膛口波在枪声检测与定位中的作用，并设计了一套基于正三角形麦克风阵列的枪声定位与测距软硬件系统。

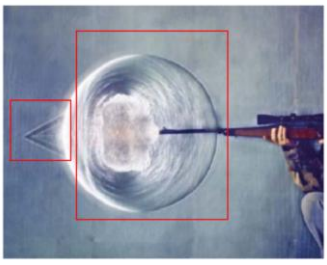


图 1.21

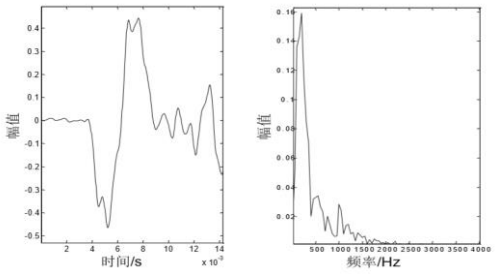


图 1.22

对于声信号处理和声学特征的研究，赵力等[4]给出了常用的信号加窗成帧、端点检测以及常用声学特征的计算方法；韩纪庆等[5]对声学事件检测技术与常用模型做了综述性介绍。徐大为[6]等对比了基于不同声信号特征的端点检测方法，并分析了他们对噪声的抵制能力和运算实时性。

针对枪声的信号处理与声学事件检测研究中，蒋小为和张文等[7]通过低通滤

波和谱减法针对膛口波进行去噪处理，在实验中得到了与理论波形高度相近的膛口波信号波形，如图 1.22 所示，并提出可以使用相关分析进行枪声检测。张克刚[8]等人研究了基于短时能量分析对枪声信号进行端点检测的方法，并提出使用持续时间处理来剔除瞬时大能量噪声。张涛[9]、张文[10]、朱强强[11]等人的研究中指出，可以采用 MFCC 作为目标片段的特征，用于进一步给分类器进行分类检测。

对于声学事件检测的分类器，Clavel 等[12]讨论了监控环境中的枪声检测，并通过 PCA 选择 13 维特征作为 GMM 模型的输入特征；刘力维[13]等提出使用 10 阶中值滤波处理端点检测中的能量序列，并用 GMM 对目标片段的按 MFCC 特征进行分类；朱强强[13]分析了 Logo、FFS、Adaboost 三种特征选择算法，用特征选择算法对时域特征、频域特征、感知域特征、基于自相关函数的特征等共计 9 个特征组成的特征全集进行特征选择，并最后输入到 GMM 中进行分类；Pimentel[14]等人提出了通过分析聚类过程中的 WSS 指标来确定聚类算法中聚类中心数目的方法。

关于声学事件数据库，Fonseca[15]等人所在的庞培法布拉大学（Universitat Pompeu Fabra, Barcelona）音乐技术研究小组为了解决目前数据驱动型（data-driven）声学计算研究所遇到的瓶颈和困难，发起了 Freesound Datasets 项目，并建立了一个基于众包（crowdsourcing）、规模宏大、音频种类较齐全的大型公开数据库 Freesound[15][24]；坦佩雷理工大学(Tampere University of Technology，TUT)信号处理学系 Mesaro 等人[16]发起了事件检测挑战 TUT Sound Events Challenge 与声学场景检测挑战 Acoustic Scene Classification Challenge，加速了基于声学的事件检测和场景分析的相关研究。

1.3 课题工作

本课题结合系统设计要求，选用 VM2020 麦克风、AD7865-1 模数转换模块、SRAM 拓展 IS61LV25616AL、FLASH 拓展 SST39VF800、AMS1117 电平转换芯片、TMS320F2812 数字信号处理器设计了如图 1.31 所示的声学检测算法所依赖的硬件系统；结合突发公共安全事件和室外道路街道场景，如图 1.32，设计了从滤波降噪、端点检测、特征提取到分类器分类的成套软件解决方案，并基于已有数据集进行了算法验证。

图 1.31 硬件系统设计框图

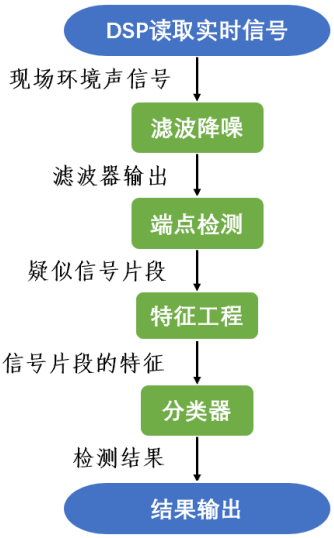


图 1.32 算法流程设计

论文的内容和章节安排如下：

第一章：介绍了本课题的选题背景和使用声学手段解决室外公共场所突发安全事件检测的原因，以及目前已有的声学检测方法的研究现状。

第二章：结合系统设计要求，分析了各类型常用的传声器器件，并确定了适用于本系统设计的具体传声器选型。

第三章：综合性能指标和成本因素，确定了 ADC 和 DSP 的选型，简要介绍了供电模块的选用以及 DSP 的 SRAM 和 FLASH 拓展方案，完成了声学检测算法所依赖的整体硬件系统的设计；

第四章：研究了包括滤波去噪和端点检测两部分在内的前端信号处理算法，对不同的解决方案进行了简要介绍。确定了基于 Butterworth 滤波器的低通滤波去噪方案和基于短时能量分析与持续时间分析的端点检测算法，并使用 MATLAB 进行了算法验证。

第五章：研究了基于 MFCC 的音频片段特征化方法，简要介绍并对比了不同分类器方案在本应用中的优劣点，确立了基于 GMM 的分类器设计方案。引入了 Elbow 方法确定聚类核心的数量，并用其设定 GMM 的高斯核数目。介绍了训练 GMM 所需的 K Means 初始化和 EM 聚类算法。最后使用 MATLAB 对各算法进行了仿真验证。

第六章：针对报告时听众提出的问题进行了回答和解释，并在最后对进一步的研究方案及其设计可行性进行了简要的讨论。

第二章．传声器选型

传声器的选型是硬件选型的重点，其稳定性、收音效果将直接影响检测的结果，因此将着重介绍选型指标、具体的参数。

2.1 选型要求

由于检测的是低频信号，因此低频性能好（放大、不失真）；需要大面积使用，价格不能过高；能耗尽量低；收音范围合适；在外界复杂环境中使用，必须受温湿度影响尽可能小；体积不能特别大；产品的质量尽量高、使用寿命尽量长、安装和维修成本低；承受声压尽可能大，满足使用需求；收录声压较高、脉冲较大的声源必须使用较低灵敏度麦克风。

由系统选型的要求可以得出选型参数的优先原则：必要参数（最大声压级（AOP）、频率响应、瞬时响应）是否达标>稳定性>价格>其他性能参数,具体相关参数会在 2.2 节中详细介绍。

2.2 指标概述

传声器指标主要包含了技术指标、市场指标、声学指标这三个方面。其中技术指标与市场指标是系统中需要重要考虑的地方，而声学指标主要是指在某些特定频率点其收录的特性，主要是针对人声录制等领域，不在系统的设计范围内。

麦克风的灵敏度是指其输出端对于给定标准声学输入的电气响应。用于麦克风灵敏度测量的标准参考输入信号为 94dB 声压级或 1Pa 的 1kHz 正弦波。对于固定的声学输入，灵敏度值高的麦克风比灵敏度值低的麦克风输出的电信号幅度高。麦克风灵敏度通常是负值，因此，灵敏度越高，其绝对值越小。模拟麦克风的灵敏度通常用 dBV 来规定，即相对于 1.0Vrms 的比值（dB）。数字麦克风的灵敏度通常用 dBFS 来规定， 即相对于满量程数字输出（FS）的比值（dB）。 基于爆炸、枪声等声压较高、脉冲较大的声源应选用低灵敏度传声器。下面列举了两种灵敏度的具体计算公式：

$$Sensitivity_{dBV} = 20 \times \log_{10}\left(\frac{Sensitivity_{mV/Pa}}{Output_{REF}}\right)$$

$$Sensitivity_{dBFS} = 20 \times \log_{10}\left(\frac{Sensitivity_{\%FS}}{Output_{REF}}\right)$$

方向性描述麦克风的灵敏度随声源空间位置的改变而变化的模式。由于不需要定向收录，故选用全指向型。图 2.31 中列举了所有的方向型。

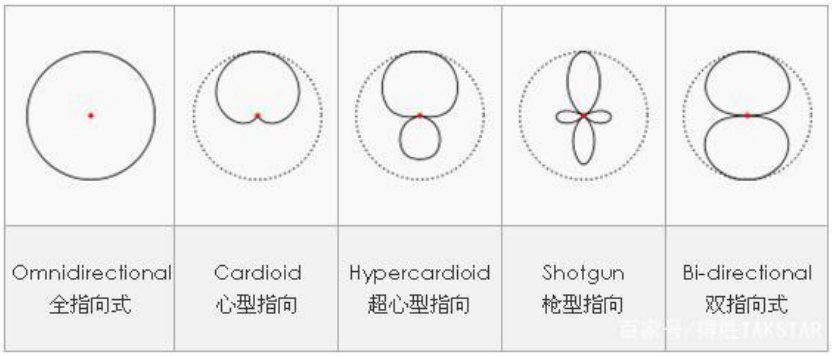


图 2.31 麦克风的指向性

信噪比（SNR）表示参考信号与麦克风输出的噪声水平的比值。SNR 为噪声水平与标准 1kHz、94dB SPL 参考信号的 dB 差。信噪比越高越好。这个参数主要影响音质，本次设计没有特别高的要求，但是同样越高越好。

麦克风的频率响应描述其在整个频谱上的输出水平。低频的频响是这个设计的重要指标之一。

总谐波失真（THD） 衡量在给定纯单音输入信号下输出信号的失真水平，用百分比表示。此百分比为基频以上所有谐波频率的功率之和与基频信号音功率的比值。THD 值越高，说明麦克风输出中存在的谐波水平越高。根据经验，输入水平每提高 10dB，THD 会提高 3 倍。高输入水平下的 THD 是这个设计的重要指标之一，越低越好。

电源抑制比是表征麦克风输出对于电源输入噪声抑制能力的参数。这个不是主要要求，但仍然越高越好。

麦克风的阻抗是具有麦克风电阻、电感和电容的电路里，对电路中的电流所起的阻碍作用称作麦克风阻抗。因为即使阻抗不匹配，也能正常使用，所以不是重要的参数。

最大声压级（AOP）指的是麦克风输出 THD 等于 10%时输入的声压大小（SPL），通常也称为麦克风的削波点。声压高于 AOP 的输入会造成输出信号严重失真。这是本次设计的重点指标之一，爆炸枪声都是高声压输入，因此 AOP 必须足够大。

一致性表示麦克风在焊接后能否保持原有性能的指标，一致性较好是麦克风性能保障的重要参数。

瞬时响应即麦克风对瞬态输入的电学反应。这个指标是这次设计中重要的指标，瞬时响应不好就无法测量。

基于本次设计的目的，必须有较广的覆盖面，即大的数量来实现，故产品的价格不能过高；更小的能耗也是一部分使用成本的体现，但不是关键要求；产品必须要具备良好的稳定性，这是市场指标中首要考虑的因素；在性能不下降的情况下使用寿命越久越好，节约人力及产品成本；所选用型号应该有较强供货能力，且持续生产。

2.3 需求转化

在常温下球面声波随距离衰减的表达式为：

$$L_p = L_w - 20lg(r) - k$$

式中L_p为声压级，L_w为基准声压级，r 为到声源的距离，k 为修正系数，自由空间 k=11，半自由空间 k=8。

距离$r1$和$r2$之间的声压级差值为：

$$L_{p1} - L_{p2} = 20lg(\frac{r2}{r1})$$

当$r2/\ r1$=2 时，衰减 6dB,也即距离加倍声压级衰减 6dB。

枪声在 1m 处声压级在 130-155dB 之间，根据声压的距离衰减公式每增加一倍距离衰减 6dB，8m 处大约在 106-131dB，因此对传声器 AOP 要求至少在 131 以上。其次枪声爆炸等都是瞬时声波，需要瞬时响应性能好；对低频要求敏感，所以选用低灵敏度，大振膜传声器且无变压器输出；在低频段范围内频响较好；全指向与一致性好；价格尽量中低、稳定性要求高、能耗尽可能低、使用寿命有保障、供货能力强。

2.4 种类选型

传声器根据声电转换分类可分为电动式（动圈式、铝带式），电容式（ECM、MEMS）、压电式（晶体式、陶瓷式、MEMS）、碳粒式、激光式、光纤式、矢量麦克风。

铝带式通过声音扰动磁场中金属带，通过电磁感应在金属带两端产生电压变化。其音质效果好、双向响应效果好、瞬态响应好但价格昂贵且铝片易受损伤、维修成本高、高声压会造成损坏。故不考虑选用。

动圈式通过声音通过空气使震膜振动，然后在震膜上的电磁线圈绕组和环绕在动圈麦头的磁铁形成磁力场切割，形成微弱的波动电流。电流输送到扩音器，再以相反的过程把波动电流变成声音。其简单紧固、易于小型化、不需要额外供电、不易过载（失真）、指向性好但频响和瞬态响应不够好。故不考虑选用。

电容式通过声音扰动改变金属膜板与背板的距离，引起膜板和背板间电容 C 的变化，电容器上存储的电荷 Q 也会随着变化，进而在电阻 R 上产生电压的变化，由此完成声音信号到电信号的转换。其频响特性与瞬态响应好，但价格较高、需要外部供电、且受湿度影响。

驻极体式（ECM）类似电容式麦克风，通过声音影响金属隔膜与背板距离，电容器上的电荷变化在电阻上产生电压的变化，由此完成声音信号到电信号的转换。ECM 的金属隔膜是永久性的含电荷材料，因此在使用中不必需要额外的偏置电源。其结构简单，体积小，价格低，瞬态性能好、频响特性好，但受湿度影响大、一致性差、内部可能过载（失真）、且灵敏度高。

压电式利用某些材料的压电效应(即声音造成材料的变形)产生电压的变化。其输出电平高、价格低，但频率响应较差、稳定性差。

MEMS 式分为电容式与压电式，利用各自原理集成到硅芯片上。其体积小、可SMT、产品稳定性好、不怕温湿度变化、一致性好，但价格较高。

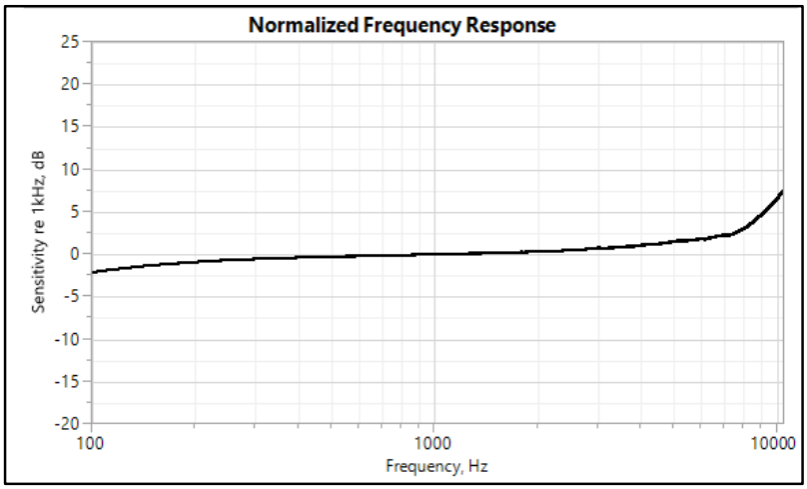
其余麦克风如矢量麦克风等用于实验室，碳粒式等已经少有人使用，主要可供选择的麦克风主要有电容式、驻极体、压电式、MEMS 等这几类。虽然较早关于资料中其他针对枪声收录的麦克风选型有选择驻极体麦克风的，经过比较其他技术的综合，并结合当前的技术实际，选用了 MEMS 压电式麦克风。其优点有信噪比高、受湿度、尘土、温度影响小、一致性好、支持单端与差分输出、电源抑制比比传统的高 30dB、声学过载点可以达到 150dB 的最大声压级。缺点仍然十分明显，就是价格高，且瞬时响应与低频频响比驻极体稍差综合考虑下，在一个比较长的周期内，由于可能出现的损坏、维修、安装、使用寿命等其他因素，实际成本较驻极体式并没有那么高。因此最终选择了 Vesper 公司的 VM2020 作为最终的传声器选型。

其特性有超高声学过载点；差分模拟输出；零件间差异小；耐用的压电 MEMS 构造。其价格为 2.6 美元。具体的参数如下图 2.41 所示。

Parameter	Symbol	Conditions	Min.	Typ.	Max.	Units
Acoustic Specifications						
Sensitivity		1 kHz, 94 dB SPL	-66	-63	-60	dBV
Signal-to-Noise Ratio	SNR	94 dB SPL at 1 kHz signal, 20Hz to 20kHz, A-weighted Noise		50		dB(A)
Total Harmonic Distortion	THD	94 dB SPL		0.1		%
Total Harmonic Distortion	THD	149 dB SPL		1		%
Acoustic Overload Point	AOP	10.0% THD		152		dB SPL
Roll Off Frequency		-3dB at 1KHz			80	Hz
Directivity			Omni			
Polarity		Increase in sound pressure	Increase in output voltage			
Electrical Specifications						
Supply Voltage			1.6	1.8	3.6	V
Supply Current		V_{Supply} ≤ 3.6 V		248		µA
Power Supply Rejection Ratio	PSRR	VDD = 1.8, 1kHz, 200mV_{PP} Sine wave		90		dB
Power Supply Rejection	PSR	VDD = 1.8, 217Hz, 100mV_{PP} square wave, 20 Hz – 20kHz, A-weighted		-112		dB(A)
Output Impedance	Z_{OUT}			1100		Ω
Output DC Offset		Both Vout+ and Vout-		0.8		V
Startup Time		Within ±0.5dB of actual sensitivity		200		µS

图 2.41 VM2020 参数表

其具体的频响曲线如下图 2.42，可以看出整体在低频段较为平直，频响较好



Normalized Frequency Response

图 2.42 VM2020 频响曲线

第三章．配套硬件系统设计

与第二章介绍的传声器相匹配的，还需要一套相应的处理与转换系统，本章将详细介绍合适的 DSP、ADC、存储扩展与电源模块的选型设计及相关参数。

3.1 DSP 选型

DSP 的选型主要是基于成本考虑的，即在保证运算与稳定性的同时最大限度地降低成本，本章中将介绍 DSP 的选型要求、指标与具体参数。

DSP 的选型要求主要包括三个方面，即精度满足要求、处理速度满足要求以及足够的外设资源。

DSP 按数据格式分为定点式与浮点式。定点式优点有体积小、功耗低、价格低、接口多、结构简单而浮点式优点有运算精度高、动态范围大、地址总线宽（寻址空间广）。综合成本和性能考虑，最终选择定点式 DSP。

图 3. 11 TMS320F2812 芯片

TMS320F2812，定点 32 位，性价比高（几十人民币），处理性能可达 150MIPS，IO 口丰富，有两个串口、两个独立的采样保持电路，哈佛总线结构，快速中断响应。片内 128k*16 位的片内 FLASH，18k*16 位的 SRAM，4M 线性程序与数据寻址空间。符合了检测系统运算的需求。其参数表如图 3.12 所示

FEATURE		TYPE(2)	F2810	F2811	F2812
Instruction Cycle (at 150 MHz)		–	6.67 ns	6.67 ns	6.67 ns
Single-Access RAM (SARAM) (16-bit word)		–	18K	18K	18K
3.3-V On-Chip Flash (16-bit word)		–	64K	128K	128K
Code Security for On-Chip Flash/SARAM/OTP		–	Yes	Yes	Yes
Boot ROM		–	Yes	Yes	Yes
OTP ROM (1K x 16)		–	Yes	Yes	Yes
External Memory Interface		0	–	–	Yes
Event Managers A and B (EVA and EVB)		–	EVA, EVB	EVA, EVB	EVA, EVB
• General-Purpose (GP) Timers		–	4	4	4
• Compare (CMP)/PWM		0	16	16	16
• Capture (CAP)/QEP Channels		0	6/2	6/2	6/2
Watchdog Timer		–	Yes	Yes	Yes
12-Bit ADC		0	Yes	Yes	Yes
• Channels			16	16	16
32-Bit CPU Timers		–	3	3	3
Serial Peripheral Interface (SPI)		0	Yes	Yes	Yes
Serial Communications Interfaces A and B (SCIA and SCIB)		0	SCIA, SCIB	SCIA, SCIB	SCIA, SCIB
Controller Area Network (CAN)		0	Yes	Yes	Yes
Multichannel Buffered Serial Port (McBSP)		0	Yes	Yes	Yes
Digital I/O Pins (Shared)		–	56	56	56
External Interrupts		–	3	3	3
Supply Voltage		–	1.8-V Core (135 MHz), 1.9-V Core (150 MHz), 3.3-V I/O		
Packaging	128-pin PBK	–	Yes	Yes	–
	176-pin PGF		–	–	Yes
	179-ball GHH		–	–	Yes
	179-ball ZHH		–	–	Yes
Temperature Options	A: –40°C to 85°C	–	Yes	Yes	Yes
	S: –40°C to 125°C	–	Yes	Yes	Yes
	Q: –40°C to 125°C (AEC-Q100 Qualification)	–	Yes	Yes	PGF only

图 3.12 TMS320F2812 参数表

3.2 ADC 选型

ADC 选型主要以保证精度与速度为前提，选用较低成本的 A/D 转换模块。

低速A/D	高速A/D
线性误差	线性误差
微分误差	微分误差
电源电流	电源电流
功 耗	功 耗
转换时间	转换率
失调误差	失调误差
增益误差	增益误差
	信 噪 比
	信噪失真比
	无杂散动态范围
	总谐波失真
	二次谐波
	三次谐波

图 3.21 ADC 性能指标

图 3. 22 AD7865-1 模块

综合图 3.21 的性能指标，最终选用了如图 3.22 所示的 AD7865-1 模块。其特点有转换高速、4 通道、14 位、采集速度 350ksps、高输入范围（0-10V）、低功耗且价格相对较低。

3.3 存储拓展与电源设计

由于分类器参数所占空间较大因此选用一片 SRAM（IS61LV25616AL），选用两片 FLASH（SST39VF800）作为片外扩展满足实际需求。

供电采用 AMS1117 芯片将 5V 转化为 3.3V IO 口电压及 1.9V MIC 及内核电压使系统正常运行。

第四章．信号处理前端

声学事件检测系统的信号处理前端负责对所获取的声信号进行信号处理，并对目标信号片段进行提取。本章将结合 TUT Acoustic Scene[8]和 Freesound[15] [24]两个声学事件数据库的部分信号样本，介绍信号处理前端的两个环节：滤波降噪和端点检测，以及对应的 MATLAB 算法仿真结果。

4.1 滤波降噪

为了对所采集的声信号进行初步的去噪处理，以减小噪声信号对目标信号片段的提取和分类的干扰，信号处理前端设置滤波降噪环节。为设计合适的滤波器，首先需要对目标信号片段的特性进行分析，从图 4.11、4.12、4.13 可以看到，枪声、爆炸声和汽车喇叭声信号的主要成分都在 1000Hz 以下。取 1000Hz 作为低通滤波的截止频率，能够去除高频噪声，同时不会损失目标信号的主要成分，不会使目标信号失真。

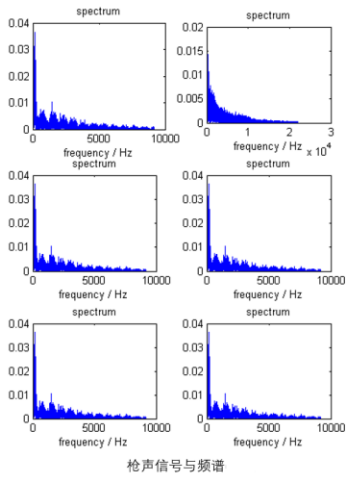


图 4.11 枪声

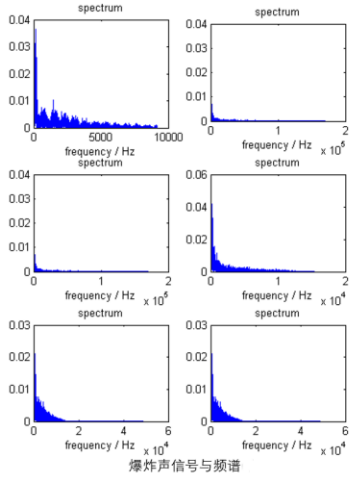


图 4.12 爆炸声

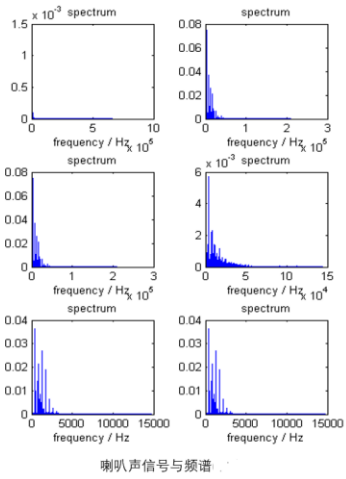


图 4.13 汽车喇叭声

对于滤波器的具体设计，采用 3 阶 Butterworth 滤波器进行低通滤波，并取截止频率为 1000Hz, 如图 4.14 所示。

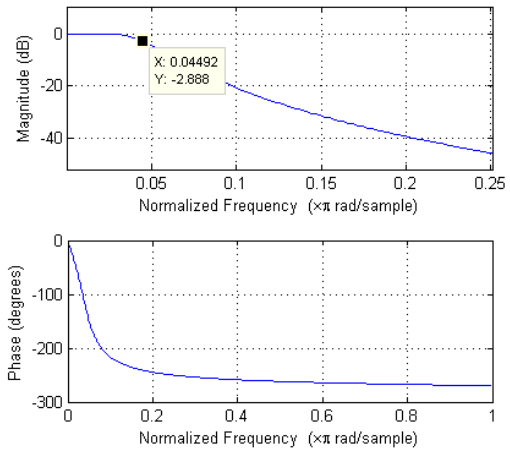


图 4.14 3 阶 Butterworth 滤波器

通过低通滤波去除高频噪声后，目标信号的大致波形往往能显现出来。为了进一步获取目标信号的波形轮廓以便识别，采用高阶均值滤波进行处理。如图4.15 所示，目标信号依次通过 Butterworth 低通滤波和 1000 阶均值滤波后，枪声信号的信号特征已经非常明显。

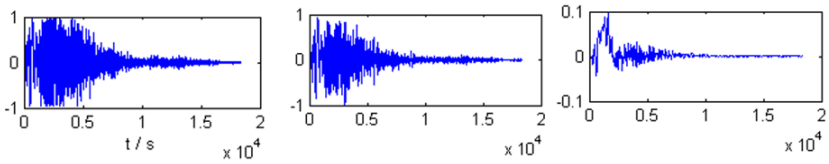


图 4.15 滤波后信号特征

对于相对纯净的信号片段，通过滤波、谱减法（Spectral Substraction）等前端信号处理手段的确能够获取目标信号的波形特征。尤其是通过低通滤波与谱减法相组合的信号处理方法，对于单发、背景噪声小、现场干扰少的情景，能够有效的获取枪声信号的波形。若再做进一步的波形分析、相关接收或是匹配滤波等基于波形的检测，一定程度上也能够实现目标信号的检测识别。对于靶场、测试基地、郊外等声源少、背景噪声小、干扰少的场景而言，这是一种可行的解决方案[7]。然而，考虑到本系统的应用场景公共场所，对于现场信号而言，相对纯净的信号片段是几乎不可能得到的，相反，现场传声器所采集到的数据必然是不同类型声音混叠、背景嘈杂，甚至会有多个或多种目标信号并发的情况。另外，考虑到公共场所的场景，尤其是室内和街道场景，声音传播的多径效应明显，对于大功率的目标信号，这个问题更为显著。无论是多声源、多目标还是多径效应，都必然会造成波形的混叠。多个波形一旦混叠，基于波形的检测便无从谈起。

结合上述分析，传统的基于波形的检测方法的应用场景与本系统的应用场景有较大差异，因此不采用基于波形的检测方法，而是通过目标信号的特性，将目标信号从背景中分离出来，再提取信号片段的声学特征，通过分类器对声信号进行检测。

4.2 端点检测

在 4.1 节中已经提到，本系统的声学检测算法首先分析目标信号的特点，按其特点将目标信号从背景中分离，再提取该信号片段的特征，最后使用分类器对声信号进行分类，从而达到对目标信号的检测。而将目标信号从背景中分离，可以转化为一个端点检测问题[18]。本系统所针对的突发公共安全事件的声信号（枪声、爆炸等）均为大功率信号，前背景分离也就简化为在现场信号中，确定大功率信号端点的问题。

端点检测（Endpoint Detection）是指，从一段声音信号中准确找出目标信号的起始点和结束点，它的目的是为了使有效的目标信号和无用的背景声信号得以分离。端点检测的方法大体分为两类，一类是基于阈值的方法，该方法根据声信号和背景声的不同特征，把这些特征与设定的阈值进行比较，从而达到目标声音

断点检测的目的。这种方法原理简单，运算方便，被广泛使用；另一类方法是基于模式识别的方法，需要估计目标信号和背景声的模型参数进行检测。基于模式识别的方法复杂度高，运算量大，往往很难被应用到现场实时声信号检测中[4]。因此，本系统采用基于阈值的端点检测方法。

在基于阈值的端点检测方法中，有两类最简单、运算复杂度最小、使用最广泛的方法：短时过零率分析和短时能量分析。短时分析贯穿了声学信号分析的全过程，这是因为有很多声学信号，从整体来看，其特性及表征其本质特征的参数均是随时间而变化的，因此它是一个非平稳过程，不能用处理平稳信号的数字信号处理技术对其进行分析出来。但是，虽然这些信号具有时变特性，但是在一个短时间范围内，其特性基本保持不变，即相对稳定，因而可以看作是一个准稳态过程，即具有“短时平稳性”。所以很多声学信号的分析和处理，往往建立在“短时”的基础上，即进行“短时分析”，将信号分为一段一段分析其特征参数。其中每一段称为一帧（frame）[4]。

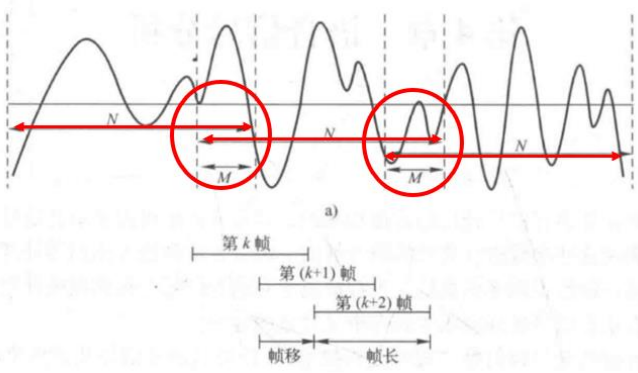


图 4.21 分帧

分帧（framing）可以采用连续分段的方法，但一般都采用如图 X 所示的交叠分段的方法，这是为了使帧与帧之间平滑过渡，保持其连续性。后一帧与前一帧的相对移动称为帧移。帧移与帧长的壁纸一般去 0~0.5。分帧使用可移动的有限长度窗口进行加权的方法实现的，即用一定的窗函数$w(n)$来乘$s(n)$，从而形成加窗信号：

$$s_w(n) = s(n)w(n)$$

在声学信号数字处理中，常用的窗函数包括矩形窗、Hamming 窗，他们的表达式如下：

矩形窗：

$$w(n) = \begin{cases} 1, 0 \le n \le (N-1) \\ 0, n < 0, (N-1) < n \end{cases}$$

Hamming 窗：

$$w(n) = \begin{cases} 0.54 - 0.46cos[\dfrac{2\pi n}{N-1}], 0 \le n \le (N-1) \\ 0, n < 0, (N-1) < n \end{cases}$$

取帧长 300，帧移 100，Hamming 窗进行分帧加窗[8]，信号被分成很多个帧进行短时分析。需要特别指出的是，分帧加窗本质是利用了信号的短时平稳性，也就是说，一帧的长度相当的短，以至于信号在这一帧内是平稳的，其特性可以认为是不变的。相应的，其能量特征在帧内的各处没有显著变化，即一帧内，信号的能量分布近乎均匀，而不是像分帧前的整个信号（比如枪声信号）存在能量集中分布。

过零率（Zero Crossing Rate，ZCR）标识一帧信号波形穿过横轴（零电平）的次数。对于连续信号，过零即意味着时域波形通过时间轴；而对于离散信号，如果相邻的采样值改变符号则称过零。因此，过零率就是样本改变符号的次数。声信号$x_n(m)$的短时过零率Z_n为[4]：

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1}|sgn[x_n(m)] - sgn[x_n(m-1)]|$$

式中，$sgn[\cdot]$是符号函数，即：

$$sgn[x] = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

而对于短时能量分析，短时能量的定义如下：

$$E_n = \sum_{m=0}^{N-1} {x_n}^2(m)$$

图 X

图 4.22 是对部分带噪声的枪声信号进行短时能量分析的结果。显然，短时能量E_n是一个度量信号幅度值和功率变化的指标。考虑到本系统目标信号的功率特征明显，采用短时能量分析进行端点检测。

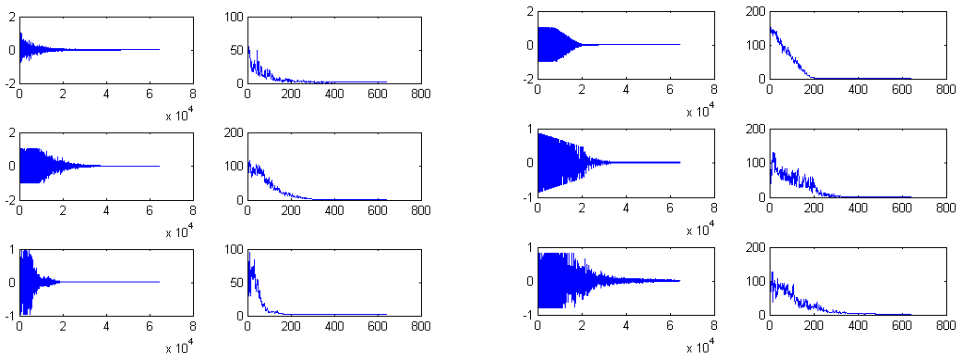


图 4.22 短时能量分析

对现场声信号进行短时能量分析后，通过合理的阈值划分，大功率的疑似信号就能从背景中被分离出来。对于阈值划分，采用自适应的短时能量阈值[8]：

$$Thr = min(E_n) + \beta[max(E_n) - min(E_n)]$$

其中，β为比例系数，E_n为帧 n 的短时能量。仿真结果指出，α取 0.4 左右时，枪声、爆炸声、汽车喇叭声三类大功率信号的端点检测准确率较高。短时能量阈

值的意义在于，对于任意帧 n，若其短时能量E_n大于阈值，则认为该帧所对应的信号片段属于疑似目标信号的一部分，该帧对应的信号片段被保留；反之，若其短时能量E_n小于阈值，则认为该帧所对应的信号片段不是疑似信号的一部分，该帧及对应的信号片段被抛弃。根据仿真结果，为了达到更稳定的阈值分割效果，可以在分割前先经过一个 50 阶的均值滤波，以提高稳定性。

在短时能量阈值分割中，有一个需要关注的问题是次要片段的影响。所谓次要片段是相对主要片段的概念。主要片段是指，产生目标信号的声学事件发生时，直接产生的声音，在传声器处采集到的的信号；而次要片段是指，在目标声学事件发生后，声音在空气中传播而出现的回响、多径等传播效应，被传声器采集到的二次信号。由于部分次要片段仍然具有较大功率，短时能量阈值分割会把这些片段当作疑似目标信号片段提取出来。但是，次要片段高度碎片化、持续时间短，并且难以作为判断声学事件是否发生的标准。因此，需要在阈值分割后将其滤除。

一种比较使用的方法是做持续时间处理[张克]。上文提到，次要片段持续时间短、高度碎片化，与主要片段相比，如果用合适的阈值对持续时间做分割，便能区分主要片段和次要片段。

$$a = \begin{cases} 0, T < t_0 \\ 1, T > t_0 \end{cases}$$

其中，a为经过阈值处理的能量的逻辑值，T 为 a 的持续正值时间，t_0为设定的持续时间阈值。仿真结果表明，针对已有的枪声、爆炸声、汽车喇叭声，取 30 个采样点做持续时间处理，能实现较好的次要片段滤除。如图 4.23 所示，(a)(b)(c)依次为部分枪声、爆炸声、汽车喇叭声的端点检测仿真结果。上方 6 个子图，蓝色为原始信号，红色为端点检测结果；下方 6 个子图，蓝色为短时能量分析结果，黑色为均值滤波结果，绿色为短时能量分割阈值，红色为持续时间分析后的端点检测结果。

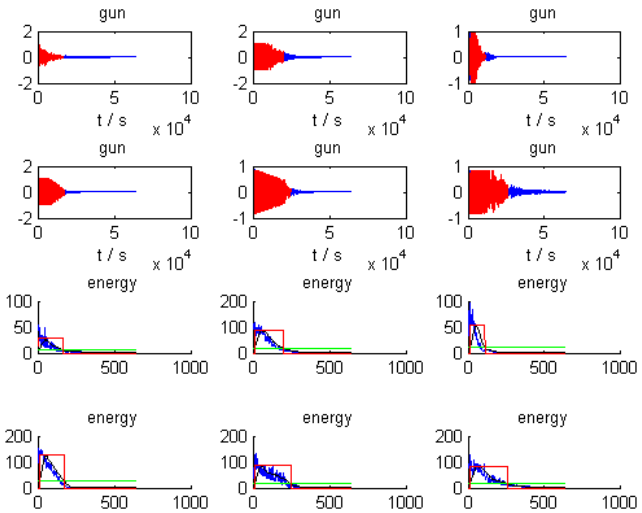


图 4.23(a)

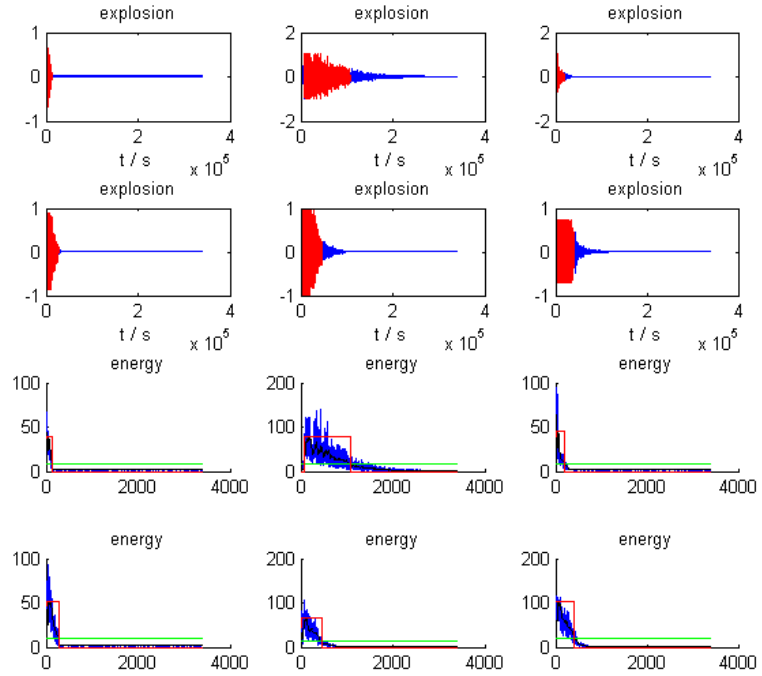


图 4.23(b)

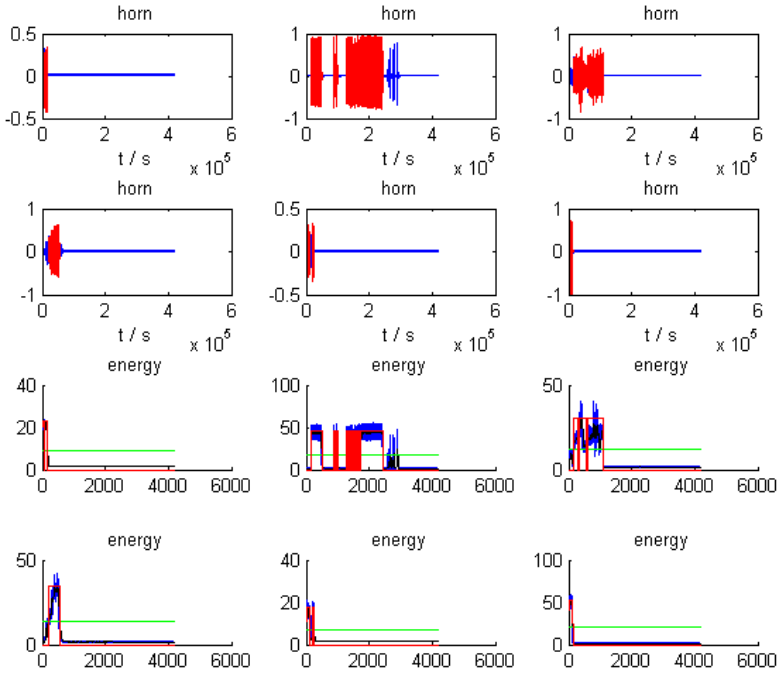


图 4.23(c)

第五章．特征工程与分类器设计

5.1 特征工程

在 4.1、4.2 节中，已经能够提取出疑似的目标信号，这些疑似的目标信号成为待检信号。将待检信号输入分类器做检测之前，需要先对其特征化。

心理学研究发现，针对不同的频率，人耳会存在不同的敏感度，因此线性的频谱并不能很好的反映人耳听觉的特性，其次由于掩蔽效应的存在，使得不同频率之间也不是相互无关的。由此人们提出了结合听觉特性的感知域的特征。这里我们主要采用 Mel 频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）特征[17]。

研究表明，人耳在低频上的感知特性成线性关系，而对于高频信号的感知则近似成对数关系[17]。为了描述人耳对不同频率的特性，人们提出了可以描绘人耳感知域特性的 Mel 频率的概念，并得到从线性频率转换到 Mel 频率的表达式[11]：

$$f_{Mel} = 2595lg(1 + \frac{f_{Hz}}{700})$$

其中，f_{Mel}表示 Mel 频率，f_{Hz}表示物理频率。根据人耳的听觉特性，所听到的声音的高低与声音的频率并不成线性关系，用 Mel 频率而符合人耳的听觉特性。临界频率的贷款随着频率的变化而变化，并于 Mel 频率的增长一致，在 1000Hz 以下，大致呈线性分布，带宽为 100Hz 左右；在 1000Hz 以上呈对数增长。类似于临界频带的划分，可以将声信号频率划分为一系列三角形的滤波器序列，即 Mel 滤波器组，如图 5.11 所示。

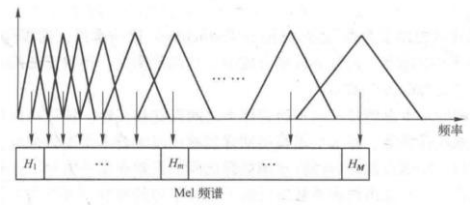


图 5.11(a)

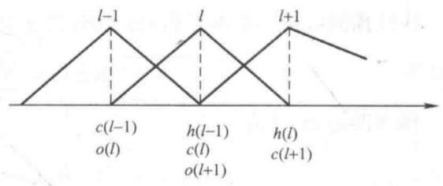


图 5.11(b)

取每个三角滤波器频率带宽内所有信号幅度加权和作为某个带通滤波器的输出，然后对所有滤波器输出做对数运算，再进一步做离散余弦变换即得到 MFCC。MFCC 参数计算过程的具体步骤如下[4]：

（1） 根据式（XXX），将实际频率尺度转换为 Mel 频率尺度。

（2） 在 Mel 频率轴上配置L个通道的三角形滤波器，L的个数由信号的截止频率决定。每一个三角滤波器的中心频率$c(l)$在 Mel 频率轴上等间隔分配，设$c(l)$、$c(l)$和$h(l)$分别是第l个三角滤波器的下限、中心和上限频率，则相邻的三角形滤波器之间的下限、中心和上限频率如图 X，并由如下关系成立：

$$c(l) = h(l-1) = o(l+1)$$

（3） 根据语音信号幅度谱$|X_n(k)|$求每一个三角形滤波器的输出

$$m(l) = \sum_{k=o(l)}^{h(l)} W_l(k)|X_n(k)| \quad l = 1,2,\dots,L$$

$$W_l(k) = \begin{cases} \dfrac{k-o(l)}{c(l)-o(l)} & o(l) \le k \le c(l) \\ \dfrac{h(l)-k}{h(l)-c(l)} & c(l) \le k \le h(l) \end{cases}$$

（4） 对所有滤波器输出做对数运算，再进一步做离散余弦变换即可得到MFCC

$$c_{MFCC}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} lg[m(l)] cos\{(l-\frac{1}{2})\frac{i\pi}{L}\}$$

此处L取 20，并选用三角形 Mel 滤波器组，对疑似目标信号片段进行 MFCC 特征的提取。图 5.12(a)是 6 个通过端点检测从背景中提取出来的汽车喇叭声疑似信号，图 5.12(b)是这 6 个疑似片段对应的 MFCC 仿真结果，不难看出其中模式的相似性。

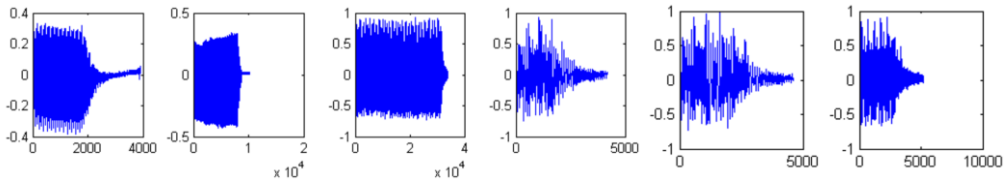


图 5.12(a) 疑似信号片段

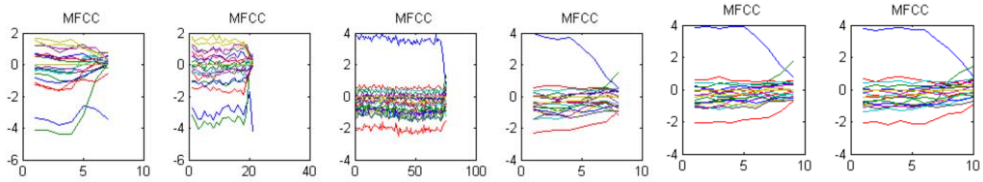


图 5.12(b) MFCC 仿真结果

5.2 分类器

经过滤波去噪和端点检测后，获得了高信噪比的目标信号前景，而经过 MFCC 特征化后，目标信号片段的信息维度被大大降低，有效的信息被进一步提取，无效的、不敏感的信息被抛弃。可以说，前面三个环节，都是在为分类器提供信噪比尽可能高、特征尽可能明显、维度尽可能合适的分类器输入。那么最后一个环节，便是设计一个合适的分类器，实现对目标信号的分类，实现对枪声、爆炸声、汽车喇叭声的声学检测。

分类（Classification）是一个根据给定数据样本进行类别预测的过程。分类的目标是，给当一个输入 x，在 K 个类别（离散值）中指定它所对应的类别 Ck。

分类往往是很多数据处理与分析的最后一步。对应的，分类器（Classifier）则是实现分类任务的数学模型。目前，在统计学习、模式识别、智能感知等领域，分类器模型及其设计获得了广泛的关注。

目前声学事件检测的主流分类模型大多基于统计学习，其中应用较广泛的包括 HMM、SVM 和 GMM[11]。HMM 对时序结构有良好的表示，其中的隐状态可以很好的刻画发音机制，使其在早期的语音识别等领域收到很大重视，也取得了很好的效果。但在枪声检测中，由于枪声发声机制相对简单，且具有突然性，另外，HMM（Hidden Markov Model）模型相对复杂，参数较多，使得 HMM 在本文所考虑的声学检测中效果并不突出；SVM（Support Vector Machine）是目前应用最广泛的分类器之一，基于判别式将样本空间进行划分。然而，一方面，在线性非线性、核函数设计、大规模样本训练等需要有较多考究[18]，另一方面，使用单独的 SVM 并不能实现本环节所需要的多分类任务。GMM（Gaussian Mixture Model）即混合高斯模型，如图 X 所示，通过采用多个高斯分量的线性组合，可以近似任意概率密度分布，并且对数据没有特殊限制。GMM 输出的似然概率可以由下式表示：

$$p(x|\theta) = \sum_{j=1}^{M} p(x|\theta_j)\,\pi_j$$

其中，x是输入，M是高斯分量数目，π_j是高斯分量j的权重，$p(x|\theta_j)$是高斯分量j的概率密度。权重之间满足归一化条件：

$$\sum_{j=1}^{M} \pi_j = 1$$

多维高斯分布的概率密度：

$$p(x|\theta_j) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\sum_j\right|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(x-\mu_j)^T \sum_j{}^{-1}(x-\mu_j)}$$

其中，D是维数，$\sum_j$是协方差矩阵，μ_j是均值[21]。GMM 模型非常简单，只有高斯分量数目一个参数，且通过计算 WSS 能够很容易确定合适的值[14]。模型训练好后，易于在检测系统嵌入式微机终端部署，能实现一站式检测[20]。

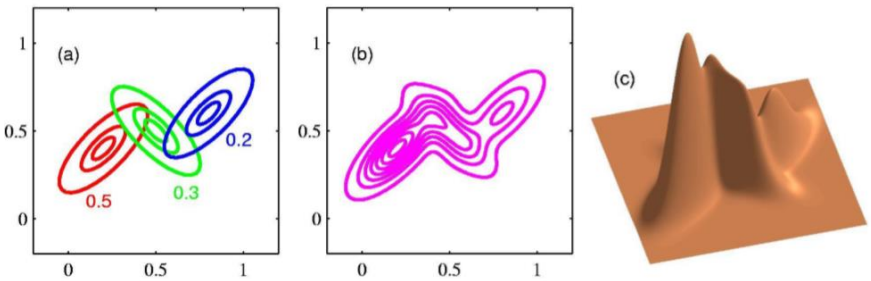


图 5.21

训练 GMM 模型的关键在于聚类（Clustering）。对于共有 N 个样本，每个样本维度 D 的数据集$\{x_1, x_2, \dots, x_N\}$，聚类的任务是将这 N 个样本划分到 K 个聚落中。一种广泛使用的算法是 K-Means 算法，分为如下几步[21]：

（1） 初始化，任意挑取 K 个聚类中心

（2） 将每个样本划归到最近的聚类中心

（3） 对各个聚落，调整聚类中心为划归到该聚落的所有样本的平均点

（4） 重复步骤（2）直到收敛

K-Means 具有实现简单、收敛性可保证等优点。但在许多 GMM 模型的训练中，K-Means 往往并不直接作为训练算法，而是用于对 EM 算法的参数做初始化[4] [21]。EM 算法（Expectation-Maximization Algorithm）是训练 GMM 的常用方法。首先，对每个样本进行软划分（soft assign），然后基于软划分重新估计每个高斯分量的权重、每个高斯分量所占样本数、每个高斯分量的高斯分布参数。如图 X 所示，迭代循环直到收敛，步骤如下：

（1） E 环节：对样本点进行软划分

$$\gamma_j(x_n) = \frac{\pi_j N\left(x_n \middle| \mu_j, \Sigma_j\right)}{\sum_{k=1}^{K} \pi_k N\left(x_n \middle| \mu_k, \Sigma_k\right)}$$

（2） M 环节：基于软划分对参数进行重新估计

$$\widehat{N_j} = \sum_{n=1}^{N} \gamma_j(x_n)$$

$$\widehat{\pi_j} = \frac{\widehat{N_j}}{N}$$

$$\widehat{\mu_j} = \frac{1}{\widehat{N_j}} \sum_{n=1}^{N} \gamma_j(x_n) x_n$$

$$\widehat{\Sigma_j} = \frac{1}{\widehat{N_j}} \sum_{n=1}^{N} \gamma_j(x_n)(x_n - \widehat{\mu_j})(x_n - \widehat{\mu_j})^T$$

GMM 通过对不同类型的声音建模，使得每种声音在特征空间中都有一个独特的区域，由此可以根据测试数据在不同模型的似然度来判断当前数据的归属。通过EM算法训练GMM简单高效，对不同的数据规模没有特别要求。同比而言，GMM 能够较好的覆盖系统设计所提出的要求。

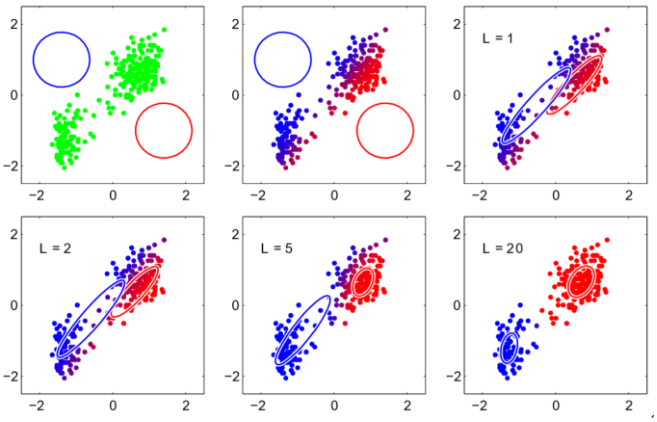


图 5.22

还有一个尚待解决的问题便是对于高斯分量数目的设计。正如前文已经介绍，高斯分量的数目本质上就是聚类中心的数目问题。直到目前，选择最优的聚类核心数目仍是聚类分析的核心研究话题之一。一种常规而典型的方法是比较不同聚类核心数目的有效性指数（validity index）。有效性指数通过聚类算法的结果计算得到，具有最优有效性指数的数目可以作为一个参数选择的方案。对于有效性指数，目前最简单而又广泛应用的是 Elbow 方法[14]。Elbow 方法是基于类内方差 WSS（Within-cluster Sum of Square）进行有效性指数的计算，WSS 衡量了聚落的紧凑程度。显然，WSS 越小，聚类效果越好。如图 5.23(a)所示，Elbow 方法的目标便是寻找一个聚类核心数目，使得新增一个聚类核心并不会再明显降低 WSS。

如图 5.23(b)所示，根据仿真结果可以发现，高斯分量数目达到 5 左右时，WSS 几乎不会再随高斯分量数目的增加而显著减小，因此取 5 作为分类器的高斯分量数目。

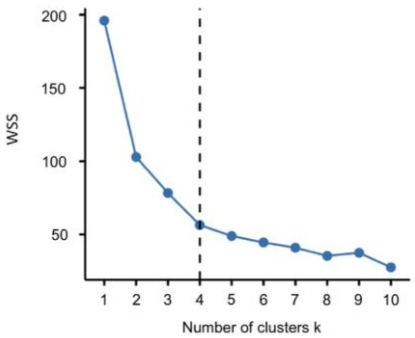


图 5.23(a)

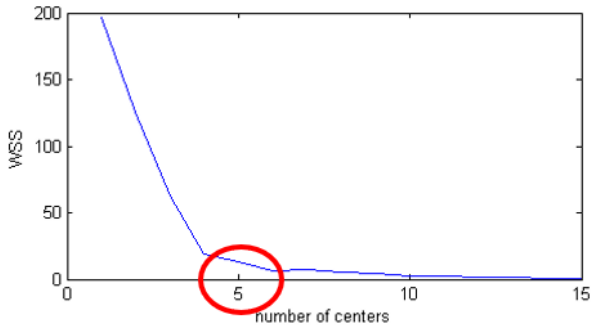


图 5.23(b)

对 3 类目标信号，分别训练 GMM。对于每类目标信号，首先通过 K-Means 初始化训练聚类参数，然后通过 EM 算法训练 GMM。由于总共有 3 个 GMM 分类器，每个 GMM 的输出都是一个描述待检信号是否属于该类别的似然概率，因此最后通过极大似然估计（Maximum Likelihood Estimation）来做分类：

$$C_n = argmax \ \ p(x_n|\theta_j)$$

基于 MATLAB 平台展开算法的仿真验证。从 TUT Acoustic Scene[8]和 Freesound[15] [24]两个声学事件数据库获取街道背景声、枪声、爆炸声、汽车喇叭声等数据，组成 60 份的小规模样本。将 80%数据作为训练集，剩余的 20%数据作为测试集。将目标信号与背景声混合后依次进行滤波去噪、端点检测、特征提取和分类器检测，训练集和测试集上正确率均达到 100%，仿真结果表明方案可行。

第六章．展望与答疑

6.1 进一步研究方向

关于“突发公共安全事件声学检测系统”的系统设计已经在前文全部介绍完毕。前文介绍的主要是一套经过算法验证、可操作性强、易于部署、成本低廉的解决方案。限于理论水平、实现复杂度和设计工程量，在鲁棒性、可拓展性、功能丰富性、场景适配性等问题上，目前的设计方案存在一些不足。本节就目前设计方案存在的问题与可优化的方案做简要介绍。

麦克风阵列是目前声学信号处理常用的手段。基于麦克风阵列可以构建波束成形（Beamforming）算法。如图 6.11 所示，波束成形器（Beamformer）是一个空间滤波器（Spatial Filter），对传感器阵列输出进行组合运算来形成特定的、有指向性的特征。换句话说，波束形成就是利用传感器阵列和阵列信号处理方法，来对特定方位的信号进行增强和对准，使得信号处理具有指向性[23]。

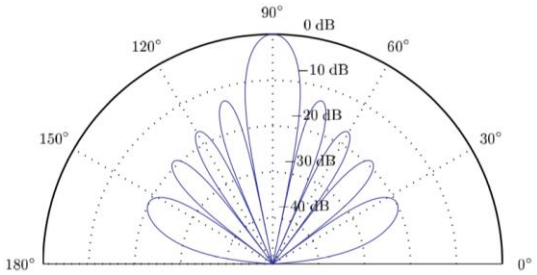


图 6.11 波束成形器

一种对于目标信号提取的优化方案在于，使用掩膜数据训练基于 BLSTM 和全连接神经网络的信噪分离器，使用这个信噪分离器，可以分别得到目标信号及其功率谱、背景信号及其功率谱，再用信、噪功率谱设计波束成形器，进一步提高输入信号的信噪比[23]。这能提高目标信号提取的正确率。

基于麦克风阵列，一个可以拓展的应用是声源定位。声源定位能够测定声学事件的发生方位，为检测系统提供更丰富的检测信息，例如并发检测的实现。基于麦克风阵列的声源定位方法包括：基于可控波束成形器的声源定位、基于到达时间差 TDOA（Time Difference Of Arrival）的定位[4]。基于可控波束的定位是最早的一种定位算法，采用波束成形方法，调整传声器阵列的对准方向，在整个接受空间内扫描，得到能量最大的方向即为声源方位。该方法是在满足极大似然准则的前提下，以搜索整个空间的方式，使传声器阵列所形成的波束能够对准信源；基于到达时间差 TDOA 的定位又称时延估计。所谓时延就是传声器阵列中不同位置单元接收到同一信源由于传输距离差异而产生的时间差。其中应用最广泛的时延估计方法是广义互相关 GCC（Generalized Cross Correlation），算法流程如图 6.12 所示。得到时延估计后再做几何解算即可得到声源位置。

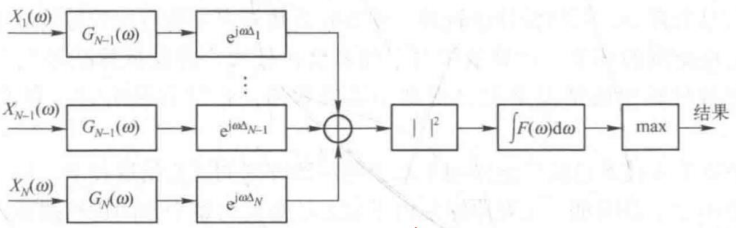


图 6.12 GCC 算法流程

本身系统的设计是从 DSP 上出发的，实际使用通信向服务器传播的只有是否接收到检测的信息以及检测到的信息种类。而随着 5G 建设逐步完善，无线通信与服务器部署的成本与方式均可以得到提升，这也是未来构建整个监控、检测、报警、记录于一体的安全系统的需要提高的部分。

针对室内检测的问题也是可以拓展的方向，但由于室内多径反射难以从根本上去除，且部署需要针对不同几何空间进行调研，这方面也是在未来需要提高的部分。

6.2 问题解答

1、枪声识别的准确还是比较重要的，建议考虑多个传声器，以及相应的多数据输入算法？

这是一个很好的建议。在多传声器设计基础上，波束成形、声源定位等更优质的检测方法和更多的声学检测功能得以开展，但这需要大幅改变原来的设计架构，多传声器所需的阵列信号处理算法也会使得设计难度大大增加。考虑到设计复杂度和理论水平深度，本文在 6.2 节中针对多传声器设计做了初步的探讨，课题主体仍以单传传声器架构作为研究对象。

2、传声器接受的原始信号很微弱，而且包含需要不同种类的噪声，如何选用放大器，保证数据来源的有效性？

选用的 VM2020 自带 BUFFER 以放大初始信号。

3、希望了解整个系统的传感器设计的设计框图。

传感部分设计思路实际上就是系统对传感器要求->对应传感器指标->根据不同传感器类型选用合适传感器类型->在该类型中选用合适传感器

4、软件算法的计算复杂度如何，可以在 DSP 芯片上进行计算吗？能否使用云计算等技术进行改进？

问题 16 对类似的算力问题做出了详细解答。

5、检测距离范围如何，在一个大型的广场上需要分布放置多少这样的检测系统？

这个要根据实际的财政预算、检测需求（需要覆盖的范围）而定，我们只是

设计系统，实际如何使用不在系统设计范围之内。

6、是否考虑过系统的经济问题，包括使用率和安装保养维护的问题？

这个在第二章中已有实际的考量，单个系统价格在 100 元以内（不包含人工安装维护费），在选用各类硬件时已经对使用寿命及稳定性有分析。

7、滤波器可不可以考虑贝塞尔滤波器，使用高阶的巴特沃斯滤波器是否会使声音失真？

首先，4.1 节滤波降噪中已经介绍，选用合适截止频率和合适阶数的 Butterworth 滤波器不会导致目标信号的主要成分造成失真。使用 Bezier 等其他低通滤波器，在不造成主要低频成分失真的前提下，也是可行的，并且对后续处理不会造成显著区别。检测的核心在于，对于提取出来的前进型号进行 MFCC 特征提取，依据 MFCC 特征进行分类。可见，只要滤波器不显著性影响目标信号的功率特性、不显著改变目标信号的 MFCC 特征，各种低通滤波其实并无显著差异，起到的主要是一个初步处理和清洗的作用。这一点其实在课堂报告中已经提到过，但限于时间没能深挖做进一步的解释，可能引起了困惑。

8、声音的“低压”是什么意思，高低压要分开考虑吗？概念的小问题。

这里的低压指的是低声压？如果是这样的话高低声压只在选择最大声压级传感器时有考虑，其次高低声压检测所需要的灵敏度也不同，这一点在第二章也有所提及。这个系统中只考虑高声压的信号，并未针对低声压进行检测。声压也会随着距离而衰减。

9、实际可行性怎么样？研讨重点都在讲系统内部结构。

实际可行性只有真正得以应用才能得知，从目前测试的情况来看，满足所述的检测要求是完全可以的。尤其是检测算法部分，使用了小规模数据（60 个样本，80%训练 20%测试）进行算法验证，细节可参见 5.2 节末尾。就目前测试情况看可行性不是问题，而且大量文献表明 MFCC+GMM 架构在声学事件检测任务中的确具有较好的性能。

10、一般公共场合较为嘈杂，怎样排除同频或过大噪音的影响？

这是一个很好的问题，虽然“同频”问题没有太理解是想表达什么意思，姑且认位是干扰信号的频率与目标信号的低频主成分接近。的确，这种干扰不会被低通滤波去除。这个问题的解答需要分 2 类：首先，如果这个干扰是小功率干扰，那么无论他是同频还是不同频，都会被端点检测部分所排除，而不会影响到后续的分类器检测；其次，如果这个所谓“同频”干扰是一个大功率干扰，那么很不幸，端点检测很可能会把它当作疑似目标信号而从背景中提取出来，送到后面的

特征化和分类器部分。解决这个问题的核心在于，尽管这个是一个“同频”干扰，但其声学特征（本系统采用了 MFCC）与目标信号存在有显著差别，在后续分类器对其进行检测时，自然能分辨出它不是目标信号。课堂报告中对 MFCC 做了详细的解释，其实只要再做进一步推理就能得到这个结论。

11、成本是否经济，算力要求是否过大？

本身这个系统就是基于低成本的设计，因此在整个系统的设计中，计算了所需的成本，每个单独的系统都在 100 人民币以下，是十分经济的。选用的 DSP 芯片被用于各类实时检测系统中，对于这个系统来说算力是足够的。MFCC 和 GMM 都是 DSP 完全能够支撑的 pipeline，对算力方面提出疑问可能是把经典分类器和深度学习混淆起来了。

12、安装位置如何考量？

我们设计的检测系统主要是针对室外检测，室外安装位置并不需要特别细致的考量，只需要房子与地面不相距太远的地方即可。而针对室内安装需要考虑的因素非常多，主要是由于多径反射问题，对不同几何空间的安装都不一样，可能需要划分几百上千个类别，需要具体情况具体分析。

13、硬件部分具体收音装置是怎样部署的，针对不同的位置和场景采用什么类型的收音设备（话筒类型）。

具体如何部署在问题 12 中已有回答，针对不同场景并不需要更换收音设备，因为传声器没有专门对某种地点与位置适用。

14、采用 GMM 分类的稳定性如何，是否采用一定量且具有代表性的测试集（真实数据）进行测试，测试的准确率如何？

不是很理解“GMM 的稳定性”，姑且认为是问训练算法的收敛性吧。5.2 节中对训练 GMM 的 K-Means 和 EM 算法做了详细的介绍，这两个聚类算法的收敛性都是可以严格证明的，证明方法有很多文献和博客可以参考，此处不再赘述。关于算法测试，第五章中已多次提到，在算法设计的时候引入了 TUT Acoustic Scene[16]权威声学数据库的数据进行算法验证，此外也用到了较流行的共享数据库 Freesound[15] [24]。限于个人电脑存储容量，没有对大规模数据进行测试，对 60 个合成样本进行了测试（20%测试 80%训练），测试集和训练集正确率均为 100%。关于这一点，课堂报告中软件架构综述就着重介绍了所引入的数据库，部分测试结果也在课堂报告的最后一部分（分类器）中给出了。

15、该检测系统是否具有一个较高的应用前景？

这一点在摘要和绪论中已经做了很多讨论了，整个系统也是针对当前一些检

测系统（尤其是安防、监控）的缺陷所设计的，他与视频手段优势互补，可应用场景还是非常广泛的。

16、信号的采集、处理已经需要较贵的硬件资源，而且在后期的信号处理需要进行机器学习和模型训练，那么更需要一个高算力的服务器，这样会进一步增加整个系统的成本，那么整个成本与系统成效是否能相互匹配？

机器学习≠深度学习，机器学习≠变态多的计算量。对于机器学习中的一部分类器，在微机终端部署训练好的模型（比如 GMM 这类）并不是很复杂或者“算力要求很高”的难题。因为模型训练好之后，在微机终端运行的可能只是计算几个公式，比如 GMM 就是用特征化的声音信号和预先训练好的模型参数计算一个似然概率，公式可以参见 5.2 节。这里或许是把机器学习中的经典分类器与深度学习、神经网络混淆了。正如 5.2 中做出了详细的说明，本系统设计使用的分类器是 GMM，不是感知机，更不是 computation-consuming 的深度神经网络，对计算设备的算力要求不高，但参数存储会有相当的开销。关于分类器参数引起的存储空间开销在硬件架构中已纳入考虑并做了适配的存储扩展，详细可以参见第三章。

17、公共场所的噪声非常复杂，如何从复杂的背景下提取出异常声音信号？

首先基于功率特征进行前背景分离，然后通过基于 MFCC 特征的分类器进行分类检测，细节可以参见第五章。

18、对于异常声音的特征表述，MFCC 只反映了声音参数的静态特性，动态特性怎么解决？

所采用的 MFCC 由连续 T 个时刻，所有 I 个余弦变换结果组成，最后输入 GMM 分类器，是具有一定动态特性的。对于动态特性很强的声音，可以使用时间-频谱图（又叫 T-F 图，Time-Frequency Spectrum），但相应的分类器可能就需要动用神经网络（CNN），设计比较复杂。就调研和算法验证来看，MFCC 已经足够了。

19、当环境中存在与枪声（或爆炸声等）频率相似但不具有危险性的声音时会不会造成错误检测？

课堂报告、问题 10、问题 17、问题 21 都在围绕一个“频率”问题，问题大同小异，都是对基于 MFCC 分类与前端去噪有相当大的误解。前端去噪只是尽其所能去除一些不必要存在的高频干扰，不是通过前端滤波来滤除全部干扰，无论干扰是否存在，最后是依靠基于 MFCC 的分类实现分类检测，只要目标信号与非目标信号的功率、MFCC 存在一定差异，原理上来说，就能区别开。

20、单点枪声识别效果较好，密集枪声是否可以很好检测？

密集枪声也能实现检测、分类、计数，考虑具体的枪声密度/射击频率，需要适当调整采样率和前背景分割阈值。

21、所检测的目标信号频率处于低频段，而检测器所身处的场景中可能也存在较多的同样处于低频段的行人噪声，是否可以做到很好的区分

该检测系统不是依据频率进行前背景分离和目标信号分类，而是基于功率特征和 MFCC 特征。低通滤波作用在于初步消除干扰而不是彻底消除干扰，低频段干扰只要其听觉特性与目标信号存在一定差异，基于 MFCC 的 GMM 分类能够区分识别。在课堂报告、4.1 节、5.4 节、问题 7、问题 10 都对这个问题做了解释。

22、并发信号对系统的影响如何处理？

目前设计主要考虑单点信号的检测。6.1 节简要介绍了基于麦克风阵列的检测方法，麦克风阵列和相应的阵列信号处理可以实现并发检测。

23、在室内与室外是否需要训练不同的模型针对不同的声学环境？

必然是需要的，这涉及到一个模型的迁移问题。针对不同的场景，必然需要用不同的数据集进行训练，比如针对街道和交通道路场景，那么训练集中的背景数据便使用街道、交通道路的场景数据；如果要应用于室内场景，那么必然用商场、旅店、餐馆等室内场景数据训练背景的 GMM，这个无需过多解释，毕竟一个系统不会同时在室外做检测同时又在室内做检测，退一步讲，即使这样一个系统需要既适配室外场景又适配室内场景，那自然需要两类场景都加入到数据集中，并且在场景切换时候需要适当的标定。需要强调的是，这并不意味着需要大幅调整系统设计，只是调整训练数据以及标定前背景分离所使用到的一些阈值参数（4.2 节）

参考文献

[1] 吴松林,杨杰,林晓东,宋波.声定位系统中的弹道波信号分析及弹道矢量计算[J].探测与控制学报,2009,31(02):54-58.

[2] 蒋灏,罗晓松,冯策,张向晖.利用 DOA 模型进行弹丸弹道轨迹估计[J].电声技术,2008(05):35-37+41.

[3] 卢慧洋. 枪声定位系统的研究与设计[D].西安科技大学,2016.

[4] 赵力. 语音信号处理. 机械工业出版社

[5] 韩纪庆. 声学事件检测技术的发展历程与研究进展[J]. 数据采集与处理, 2016,31(02):231-241

[6]徐大为,吴边,赵建伟,刘重庆.一种噪声环境下的实时语音端点检测算法[J].计算机工程与应用,2003(01):115-117.

[7] 蒋小为. 枪声信号分析与预处理[C]. 中国声学学会第十一届青年学术会议会议论文集, 2015:509-512.

[8] 张克刚,叶湘滨. 基于短时能量和小波去噪的枪声信号检测方法[J]. 电测与仪表, 2015,52(S1):130-132+138.

[9] 张涛,苏春玲.一种用于枪声的多级检测识别技术[J].电子设计工程,2013,21(18):56-58+61.

[10] 张文,赵云,马丽娜,刘继恒,陈雯,蒋小为. 固定靶场枪声信号检测和识别[C].中国声学学会.2018 年全国声学大会论文集 L 结构与建筑 M 气动声学与大气声学 N 机械振动与冲击.中国声学学会:中国声学学会,2018:12-13.

[11] 朱强强. 公共场所下的枪声检测研究[D].哈尔滨工业大学,2017.

[12] C. Clavel, T. Ehrette and G. Richard, "Events Detection for an Audio-Based Surveillance System," *2005 IEEE International Conference on Multimedia and Expo*, Amsterdam, 2005, pp. 1306-1309, doi: 10.1109/ICME.2005.1521669.

[13]刘力维,袁高高,潘志刚,董俊.基于 GMM 和枪声的军事环境判别[J].舰船电子工程,2009,29(06):103-105.

[14] Pimentel, B. A., & de Carvalho, A. C. P. L. F. (2020). A Meta-learning approach for recommending the number of clusters for clustering algorithms. *Knowledge-Based Systems*, 105682.

[15] Fonseca E , Pons J , Favory X , et al. Freesound Datasets: A Platform for the Creation of Open Audio Datasets[C]// *International Society for Music Information Retrieval Conference*. 2017.

[16] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *2016 24th European Signal Processing Conference (EUSIPCO)*, Budapest, 2016, pp. 1128-1132, doi: 10.1109/EUSIPCO.2016.7760424.

[17] Stevens, S. S., and J. Volkmann. “The Relation of Pitch to Frequency: A Revised Scale.” The American Journal of Psychology, vol. 53, no. 3, 1940, pp. 329–353. JSTOR

[18] Bernhard Scholkopf and Alexander J. Smola. 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. *MIT Press*, Cambridge, MA, USA.

[19] Burges, C.J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 121–167 (1998).
[20] A. Dufaux, L. Besacier, M. Ansorge and F. Pellandini, "Automatic sound detection and recognition for noisy environment," *2000 10th European Signal Processing Conference*, Tampere, 2000, pp. 1-4.
[21] C.M Bishop, Pattern Recognition and Machine Learning, *New* York, NY : *Springer,* 2006. - 738 p.
[22] 佘大鹏. 基于多组麦克风阵列的枪声定位算法研究[D].国防科学技术大学,2015.
[23] Jacob Benesty, Jingdong Chen,Yiteng Huang, Microphone Array Signal Processing, *Springer*
[24] https://freesound.org/