# 硕士专业学位论文

实时关键词识别的智能实体沙盘灯光控制系统的研究

Research on Intelligent Physical Sand Table Lighting Control System Based on Real-time Keyword Spotting

作者：穆彦龙

导师：李纯喜

北京交通大学

2021 年 5 月

# 学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：穆航　　导师签名：李迅

签字日期：　2021 年　5 月 31 日　　签字日期：2021 年　5 月 31 日

学校代码：10004 密级：公开

# 北京交通大学

## 硕士专业学位论文

# 实时关键词识别的智能实体沙盘灯光控制系统的研究

# Research on Intelligent Physical Sand Table Lighting Control System Based on Real-time Keyword Spotting

作者姓名：穆彦龙 学　　号：19125040

导师姓名：李纯喜 职　　称：副教授

工程硕士专业领域：电子与通信工程 学位级别：硕士

北京交通大学

2021 年 5 月

# 致谢

本论文是在我导师李纯喜老师的悉心指导下完成的。在本论文的完成过程中，从研究方向的确定到论文的选题一直到论文的最终完成，李老师一直给予了我极大的教导以及丰富的帮助。在研究方向确定时，李老师用他渊博的学识与严谨的治学态度为我指出了一个独具匠心的研究方向。在论文的选题中，李老师用他细心的育人精神与厚重的科研理论指导了我，帮助我在选题思路以及研究总体中确定了方向。在最后的论文写作中，李老师细心地帮我指出了一些逻辑上的漏洞以及实验中的不足，为我论文的完善起了很大帮助。李老师在科研教学工作中废寝忘食，经常与同学探讨问题或者修改论文直到深夜，值得我们每位同学学习。此外李老师严谨的教学态度、渊博的学识、和善的待人态度、严谨勤奋的治学风格是我学习的榜样，同时始终激励我不断进取。在此，我由衷地感谢李老师对我在学习、生活、研究等各方面无私的帮助。

同时要感谢实验室的所有老师，感谢赵永祥老师、郑宏云老师、郭宇春老师、陈一帅老师、孙强老师和张梅老师在我研究生学习期间给予我的悉心指导和无私帮助。赵老师经常在组会中提出非常独到的研究问题，能够寻找出找出问题的关键。郑老师、郭老师经常会提出许多非常巧妙的解决问题的方法，让我受益良多。陈老师在机器学习中造诣颇深，为模型的改进提供了思路。老师们不仅为我指明了研究方向，还在论文完成过程中为我提出了许多宝贵的意见，帮助我解决研究中所遇到的困难。在此我向实验室的各位老师表示真挚感谢。

在研究生学习期间和撰写论文期间，还要感谢实验室的高志朋学长、崔子琦学姐、张虎信学长、张琛玥学姐以及赵映南、李从、李文雯、郑东霞、邓金雪等同学以及我的舍友赵全城和杨子林同学对我生活和学习上的帮助，感谢你们陪我度过这段难忘的学习生涯。在每次遇到困难时，与他们的沟通交流给予了我很多建议和启发，帮助我走出困境。在此向他们表达我的感谢之意。

最后，特别感谢一直无微不至地关心、支持我的父母和其他亲人朋友，正是他们热情的鼓励和默默的奉献，才使得我顺利地完成学业，成为社会建设有帮助的人。

# 摘要

智能实体沙盘现在被广泛应用在众多领域，可以从多种角度向观看者展示声音与画面信息，使其有很好的观看体验。但是如今的智能实体沙盘依旧存在着一些不足之处，比如控制手段单一，基本上都是使用按钮或触屏来控制，依旧存在着进一步优化的空间。为了解决这一问题，本文为智能实体沙盘增加了控制手段，使用关键词识别方法来进行沙盘演示控制。

本系统设想工作在军事讲解领域，不能连接外部网络，需要在本地进行关键词识别，并且要求有很好的实时性，在对灯光设备进行控制的情景内，需要在讲解员对一个动作进行介绍时，灯光设备就能同时做出反应。

在现有研究中，有少数智能实体沙盘增加了关键词识别模块来控制智能实体沙盘。不过这些系统中的关键词识别模块依赖于外部语音识别平台，无法做到在本地识别。此外，在目前工作在本地的关键词识别模块中，多采用事先录制好的音频文件进行识别，只追求识别精确度而不考虑识别速度，算法复杂，无法做到对从外部接收到的实时语音信号进行识别，不能满足现有需求。

为此，本文设计了一个实时关键词识别的智能实体沙盘灯光控制系统。本系统通过本地的机器学习模型来获取讲解员讲解过程中说出的关键词，将其转化为指令信号发送给沙盘，沙盘的灯光设备根据这个指令信号做出对应的灯光动作，从而在满足保密性要求的前提下，实现实时的沙盘灯光控制。

本文主要贡献如下：

(1)本文针对目前智能实体沙盘控制手段不足的问题，设计并实现了一个实时关键词识别的智能实体沙盘灯光控制原型系统，通过获取讲解员在连续语句中发出的关键词信息，实现了对沙盘灯光系统的实时控制。系统包含实际使用部分与机器学习模型训练部分。

(2)本文设计了一个使用实时关键词识别来控制灯光设备的系统。实际使用部分可以将讲解员发出的实时信号进行获取并语音分片，之后在本地对其使用机器学习模型进行实时识别，并控制灯光设备做出对应的动作。

(3)本文训练了一个用于实时关键词识别的机器学习模型。机器学习模型训练部分使用语音合成平台产生的关键词音频作为训练输入，具有更大的灵活性，便于后续的训练。在机器学习模型训练中，挑选比较了多种模型，平衡了识别速度与准确率，选出了最适合本场景的模型。

**关键词：**智能实体沙盘；实时关键词识别；机器学习；灯光控制

# ABSTRACT

Intelligent physical sand tables are now widely used in many fields, which can display sound and picture information to viewers from multiple aspects, making them have a good viewing experience. However, today's smart physical sand table still has some shortcomings, such as less of control method, which is basically controlled by buttons or touch screens, and there is still room for further optimization. In order to solve this problem, this article adds a control method to the intelligent physical sand table, and uses the keyword spotting method to control the sand table demonstration.

This system is supposed to work in the field of military explanations. It cannot be connected to the external network, and needs to recognize keywords locally, requiring good real-time performance. In the context of controlling lighting equipment, it is necessary for the lighting equipment to respond at the same time when the guide introduces an action.

In the existing research, there are a few intelligent physical sand table that have added keyword spotting modules to control the intelligent physical sand table. However, the keyword spotting modules in these systems still rely on external speech recognition platforms and cannot be recognized locally. In addition, in the keyword spotting module currently working locally, pre-recorded audio files are mostly used for recognition, and only the recognition accuracy is pursued without considering the recognition speed. The algorithm is complex and cannot be used for real-time voice received from the outside, which cannot meet existing requirements.

In order to deal with this problem, this paper designs a real-time keyword recognition intelligent entity sand table lighting control system. This system uses the local machine learning model to obtain the key words spoken by the guide during the explanation process, convert them into command signals and send them to the sand table. The lighting equipment in the sand table makes corresponding light actions according to this command signal, satisfying the confidentiality requirements, real-time sand table lighting control is realized.

The main contributions of this paper are as follows:

(1) Aiming at the current lack of intelligent physical sand table control methods, this paper designs and implements a real-time keyword spotting intelligent physical sand table lighting control prototype system. By obtaining the keyword information issued by the

guide in the continuous sentence, the correct real-time control of the sand table lighting system is realized. The system includes the actual use part and the machine learning model training part.

(2) This paper designed a system that uses real-time keyword spotting to control lighting equipment. In the actual use part, the real-time signal sent by the instructor can be obtained and voice segmented, and then the machine learning model can be used to recognize it in real time locally, and the lighting equipment can be controlled to make corresponding actions.

(3) This paper trained a machine learning model for real-time keyword spotting. The machine learning model training part uses the keyword audio generated by the speech synthesis platform as the training input, which has greater flexibility and is convenient for subsequent training. In the machine learning model training, a variety of models were selected and compared, and the recognition speed and accuracy were balanced, and the model most suitable for the scene was selected.

**KEYWORDS:** Intelligent physical sand table; Real-time keyword spotting; Machine learning; Lighting control

# 目录

# 1 引言

## 1.1 研究背景和意义

智能实体沙盘在生活中具有广泛的应用，是很多行业不可或缺的工具。沙盘来源于军事上的战争模拟[1]，是一种根据实地地形，按照一定比例，能够反映现场实际情况的模型。用于指挥员在对战场内的实际地形进行研究、对地方情况进行了解、对演习计划进行制定和对作战方案进行规划等工作。随着沙盘制作工艺的进一步完善，在传统的军事领域中应用中，沙盘不仅可以直观地表示出战场各处的地形特点，也可以方便地让人观看到各处军事单位部署情况。因此沙盘可以帮助指挥人员制定决策，方便指挥人员研究地形以及演练战术。此外，沙盘在消防[2]、林业[3]、交通[4]以及气象[5]这些民用领域也有着广泛的应用。然而，随着时代的进步，传统沙盘拥有的问题也日渐明显，例如存在着设计速度缓慢、制作工艺复杂、制作成本高昂、制作流程漫长、精度较低、无法后期修改等诸多问题。为了解决这些问题，智能实体沙盘随之诞生。

智能实体沙盘依托新世纪的计算机技术，相比传统沙盘更加灵活可靠，可维护性大大增强，性能多样化、直观化、智能化的程度都大大增加了[6]。此外，智能实体沙盘融合了触控一体机操作技术、多媒体软硬件控制技术、电路自动控制技术以及触摸屏技术等技术，提高了使用者的体验，使得演示效果更加清晰直观，并且可以借助图片、视频、动画、语音、灯光等多个方式进行展示[7]。通过连接音箱与投影仪这样的多媒体设备，使用者可以在计算机主机上控制这些设备，播放声音与动画，让使用者获得更好的视听体验。此外，讲解员和听众也可以在手机或者平板电脑中的用户交互界面控制智能实体沙盘，使操作也变得更加简洁高效。因此智能实体沙盘在交通[8] [9]、教学演示[10] [12] [13]、视频处理[11]等领域都有着应用。可以让使用者直观地感受整片区域内发生的事件，可以逼真地模仿整片区域的地形、变化、各个单位的移动等信息，方便使用者进行调度与操控，使得使用者能够对整片区域有着直观和具体的认识。比如在常见的交通领域[8]，智能实体沙盘可以模拟出这个区域内的道路与车辆信息，能够模拟车辆的行驶与交通流量的控制，可以展示出各处交通摄像头的图像以及各个车辆的位置信息。用户在用户交互界面中点击沙盘中的对应区域，或者特定的可交互单位时，对应的区域会出现灯光闪烁，用户交互界面里会出现对应区域的具体信息，沙盘中的可交互单位则会做出反应，比如行驶，升降，灯光切换等行为。

## 1.2 研究现状

智能实体沙盘功能的多样化发展是现在智能实体沙盘的发展趋势。智能实体沙盘的功能来源于其应用着无线通信技术、红外识别技术、自动控制技术、图像处理、多媒体技术以及传感器技术等丰富多样的技术。因此，如何在智能实体沙盘中应用更多技术，来增加更多实用性的功能或者更加灵活的控制方式是现在很多工程技术人员追求的目标。

在国际中近几年对智能实体沙盘在不同领域中的应用都有研究，Q. Li 等人提出了一种基于电子沙盘的情境可视化系统[14]。建立了情势可视化系统的通用架构，并将态势可视化的软硬件平台应用于多指挥官的协商与决策，提高了场景展示和人机交互的能力。G. Huang 等人开发出了一个工作在智能家居中的智能庄园沙盘演示系统[15]。该系统基于沙盘实体模型建立了 3D 虚拟模型，可以有效地模拟一些智能农业，智能家居的实际场景，在智能农业系统部分使用无线网络传感器收集作物生长环境信息，在智能家庭系统部分可以实现对庄园大门，路灯等的控制。P. Frantis 等人使用增强现实设备与各种运动跟踪设备系统设计了一个工作在军事领域的虚拟沙盘系统[16]。

在国内，为了给智能实体沙盘添加更加丰富的模块，实现更加多样的功能，国内研究人员同样做了一些工作。杜一川构建了一个智能交通沙盘系统[8]，用来模拟道路上的交通检测器的工作情况，同时对小区域的交通环境进行了模拟。这是国内比较早的对智能实体沙盘的研究。之后闫保中等人将可编程控制器与智能实体沙盘结合了起来[9]，实现了在地铁智能实体沙盘系统中多媒体智能化视频播放软件的开发。马蓉与赵九思将 RFID 技术应用在了智能实体沙盘上[10]，设计出了一套智能停车场智能实体沙盘系统，并将该系统应用在了物联网教学演示中。刘浩等人对智能交通沙盘进行了开发[11]，将树莓派微型计算机嵌入到了智能实体沙盘中，用来进行视频处理，可以实现车牌识别与车辆跟踪的功能。纪显俐等人在 Unity3D 平台上将增强现实技术与智能实体沙盘系统相结合[12]，制作出了一个用来地理教学演示的智能实体沙盘系统。李莉等人将物联网沙盘系统应用在了课堂教学中[13]，使课堂情境更加生动活泼。

## 1.3 研究内容

虽然上节所述的研究人员都为智能实体沙盘系统添加了许多模块，增加了许多功能，但是如今的智能实体沙盘依旧存在着不足。不同智能实体沙盘由于应用

场景的不同，导致了其需要增加的模块或者需要实现的功能不同。而且增加的模块只能用来实现某一特定的功能，对于其他系统开发的借鉴意义较小。这点体现了智能实体沙盘的灵活性，也展现出了对智能实体沙盘系统开发的复杂性与多样性。对于一个用来讲解演示的智能实体沙盘系统，在进行演示播放时，讲解员必须按照固定的顺序对声光图像进行播放，不能根据现场实际情况进行变通，缺乏灵活性。同时控制方式不够丰富，基本上都是使用按钮或触屏来进行控制，依旧存在着进一步优化的空间。语音控制如今在众多场合内都有应用，操作简单快捷，易于学习，可以丰富智能实体沙盘的控制方式。综上所述，使用语音来控制智能实体沙盘成为了一个值得研究的课题。

为了丰富智能实体沙盘的控制方法，实现更加灵活的功能，本文设计了一个实时关键词识别的智能实体沙盘灯光控制系统。本系统的独特之处有两点，一点是本系统将本地关键词识别模块与智能实体沙盘系统相结合，改变了现在智能实体沙盘系统中关键词识别模块全都是基于网络语音识别平台的状态。另一点是本系统的关键词识别部分基于机器学习模型，识别速度快，可以满足实时性的需求。

本系统分为训练和应用两部分。在训练部分，系统从语音合成平台中获取关键词音频文件作为数据集，使用音频关键词文件训练机器学习模型。在应用部分，在讲解员进行讲解时，系统将获取到的实时语音信号进行分片并使用机器学习模型对语音分片进行关键词识别，检测其发出的特殊关键词，并转化为控制指令信号传递给 DMX512 设备。DMX512 设备将控制指令信号转化为灯光设备的通道信号，从而控制灯光设备。下面为本文的研究内容：

第一，训练部分包含音频数据获取模块与机器学习模型训练模块两个模块。音频数据获取模块用于从语音合成平台获得用于训练的指定关键词音频文件。机器学习模型训练模块用于使用从关键词音频中提取到的语音特征作为输入来对机器学习模型进行训练。在训练之前需要对关键词音频数据进行预处理与特征提取。在训练之后比较不同模型的识别速度与精确率，从中挑选出适合的模型供应用部分实际使用。

第二，应用部分包含实时关键词识别模块与灯光设备控制模块两个模块。实时关键词识别模块对输入的实时语音信号进行语音分片、时间标记、端点检测以及能量检测的操作，之后经过预处理与特征提取后将其 MFCC 特征作为输入，使用训练部分提供的机器学习模型对 MFCC 特征进行实时识别，产生每个关键词的识别概率。经过阈值判断后，如果被识别为一个关键词，便生成一个指令。之后该模块通过串口通信将指令传递给灯光设备控制模块。在灯光设备控制模块中，预先为每个关键词设置好对应的灯光动作，并保存在灯光动作库中。在实时关键词识别时，其中的 DMX512 设备通过串口通信接收到实时关键词识别模块

发送过来的指令信号后，根据指令信号从灯光动作库中读取对应的灯光通道信息，根据 DMX512 协议将指令信号转化为 DMX512 数据包。数据包内包含有用来控制灯光设备的各个通道信息。灯光设备由串口通信接收到这些通道信息后，在智能实体沙盘上做出该关键词对应动作，比如灯光的照明、移动、消失等。从而实现用关键词识别来控制灯光设备的功能。

## 1.4 文章组织结构

本论文总共包含五个章节，每个章节的主要内容如下：

第一章为引言部分。主要介绍了智能实体沙盘系统的研究背景、意义以及研究现状，之后对本文的研究内容和文章结构安排进行了介绍。

第二章为相关技术部分。主要介绍了与本系统有关的相关研究，包括对语音信号进行处理的方法以及关键词识别的发展，介绍了机器学习模型以及机器学习模型在关键词识别中应用历史的发展。最后还对 DMX512 协议进行了介绍。

第三章为系统总体设计和实现部分。首先介绍了设计本智能实体沙盘系统的原因，本系统依据的原型系统、整体的物理架构以及面临的挑战。然后介绍了智能实体沙盘系统的实现方法，对设计的基本思路以及系统的结构进行了介绍。最后详细介绍了系统包含的四个模块的具体实现方法与原理。

第四章为系统评估部分。主要介绍了实验中的软硬件环境与评价步骤，评价了系统的实时关键词识别模块以及灯光设备控制模块两个模块工作情况。

第五章为结论部分。包括对本文的工作内容总结以及对未来工作的展望。

# 2 相关技术

本章介绍了本实时关键词识别的智能实体沙盘灯光控制系统所需要的相关技术。包括语音信号的预处理与特征的提取方法，关键词识别技术的研究现状，机器学习模型的选择与用法以及 DMX512 协议的内容。

## 2.1 语音信号的处理

### 2.1.1 预处理

关键词识别的第一步是对要进行识别的音频数据进行预处理，进行完善的预处理能够使得后面的语音提取和识别更加有效。预处理由如下几步构成：采样、预滤波、预加重、分帧、加窗。采样用于获取连续的语音信号，并将连续语音信号转化成计算机主机比较容易处理的数字信号。预滤波用于保留有用的语音频率部分。因为人发声的频率范围是有一个大致分布频段的，在这个频段之外的声音就可以认为是外部噪声，不应该输入进系统。预加重则用来防止高频部分有用的信息遗失。这里的低频与高频是在人发声范围内的相对低频与高频。因为高频中通常包含更多有用的信息。对人的发声研究表明，语音信号从口唇辐射后会有 6dB/Oct 的衰减。不过这种衰减对高频部分影响更大，会使高频部分衰减得比较严重，从而导致有用的信息损失。进行预加重步骤的动机就是为了防止以上问题的出现，并提升高频部分的信噪比。预加重部分的传递函数为：

$$H(z) = 1 - az^{-1} \tag{2-1}$$

其中$a$为预加重系数，范围应该处在 0.9 到 1 之间，通常取 0.97。分帧用来获取短时平稳信号。人在实际发声的时候，发出的声音是一个随机信号，随机信号的参数随时都在变化，而且完全无法预测。不过从人发声的动作来看，肌肉的运动过程的变化相对于发出的声音的变化是相当缓慢的，在短时间内声音的变化幅度比较小。因此，可以把语音信号当作一个短时平稳信号，在很短的一个时间段内是平稳的。这平稳的时间段被截取出来之后被称为帧，帧对应的时间长度叫做帧长，帧长通常为 10~30ms。不过考虑到相邻两个帧的连接处的信号会急剧变化，为后面的加窗以及特征提取造成干扰，所以两个相邻的帧通常要有一定的重叠。信号经过分帧后会产生频谱泄漏，为了减小频谱泄漏带来的影响，需要对其进行加窗，窗的长度一般等于帧长。窗两端的坡度要小，使信号平稳过度到零，从而减小截断效应。

### 2.1.2 MFCC 特征提取

梅尔倒谱系数[17](MFCC)是在梅尔标度频率域提取出来的倒谱参数，用来描述人耳频率的非线性特性。根据 Stevens 和 Volkman 在 1940 年对人耳听觉的研究[18]，人耳实际感受到的声音频率与原始声音的实际频率在一定范围内呈对数关系，如图 2-1 所示：

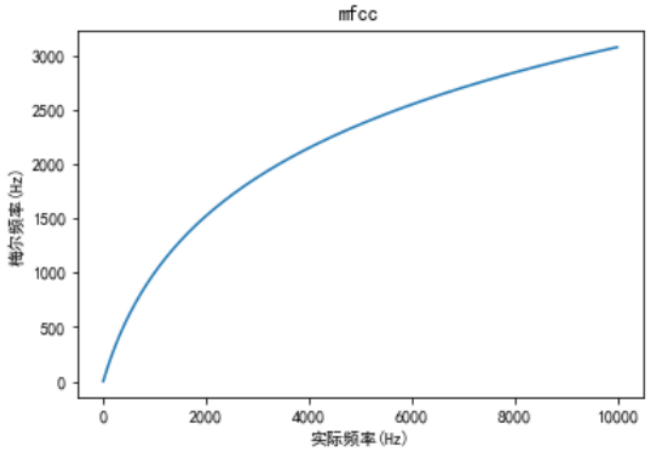


图 2-1 梅尔频率与实际频率关系图

Figure2-1 Relationship diagram between Mel frequency and actual frequency

这个关系可以近似使用式(2-2)表示：

$$F_{mel} = 2595log_{10}(1 + \frac{f}{700}) \tag{2-2}$$

其中$F_{mel}$是梅尔频率，$f$是以 Hz 为单位的实际语音频率。由于 MFCC 特征可以很好地模拟人耳的实际听觉体验，因此被广泛应用在各种语音识别模型中，被证明是在语音识别工作中最有效的特征。Schroeder 发现[19]，人类听觉系统会在对某个频率处的感知时受到周围的一个临界带的能量的影响，而且这个临界带的宽度随着频率的变化而变化。根据这一临界带，可以划分出一组滤波器，这就是梅尔滤波器组的由来。

MFCC 提取过程包括快速傅里叶变换、梅尔滤波器组、倒谱分析、离散余弦变换、动态特征提取五个步骤。快速傅里叶变换用于把离散的输入语音信号从时域变换到频域，并且输出信号的能量谱。梅尔滤波器组包含一组 M 个三角形滤波器组，用来对之前模拟人耳实际感知对通过快速傅里叶变换获得到的能量谱进行滤波。这些滤波器在低频部分密集，在高频部分稀疏，为的是模拟人耳实际对语音信号滤波时的情形。倒谱分析的目的是提取出其频谱的包络信息，方法是对频域信号进行傅里叶变换，之后再取对数。对数运算包括取绝对值和 log 运算。取绝对值的原因是仅使用幅度值作运算，忽略相位的影响，因为相位信息在语音

识别中作用不大。log 运算是为了分开包络和细节，包络代表音色，细节代表音高，语音识别关注的是其音色部分。另外，人的感知与频率的对数成正比，正好使用 log 模拟。离散余弦变换则用来分离加性的基音信息与声道信息。在同一帧频域内，基音信息变化快速，声道信息变化缓慢，做一次离散余弦变换可以将其分离，生成一个包含 25 维向量的频谱图。最后对这个频谱图进行降维，获取其的 12 维特征，合并上对数运算时产生的平均能量作为第 1 维特征，这样就得到 13 维的 MFCC 特征。

## 2.2 关键词识别技术

### 2.2.1 关键词识别

语言是人类生活中常用的交流工具，包含丰富的信息。随着计算机技术的发展，如何让机器能够理解人类的语言，提取其中令人感兴趣的信息成了一个各国研究人员的一个研究方向，语音识别技术便随之诞生了。语音识别出现之后已经给人们的生活带来了很大影响，在人机交互、在线翻译以及语音查询等多个领域中发挥着不可替代的作用。

按照识别的对象的不同来划分，语音识别可以被划分为以下三类：孤立词识别，连续语音识别和关键词识别(keyword spotting，KWS)。孤立词识别用于对单个词汇进行识别，输出单个词汇的文本信息。连续语音识别用于对连续语音信号进行识别，并转化为连续的文本信息，不过需要消耗大量的资源。关键词识别用于识别一段语音信号中的特定词汇，目标是输出这一段语音信号中包含着的系统预先设置好的关键词。在人们实际说话时，每句话中的大部分信息其实都包含在其中几个关键词中。意味着想要听懂这句话，不需要对这个句子里面的所有词语都进行理解，只需要理解其中少部分的关键词并合理安排顺序就可以做到。这个想法推动了关键词识别的发展与应用。

关键词识别系统程序需求少，处理反应迅速，广泛应用在唤醒系统、语音监听以及有限指令控制系统中。苹果手机中的“Hey siri”就是一个应用广泛的唤醒系统，在待机状态下，唤醒系统一直在接受周围的语音信号，一旦识别出用户发出的关键词语音，便立刻做出应答，唤醒系统。随着互联网的发展，关键词识别在语音监听中也发挥着作用，应用人员可以在海量的数据中筛选出他们感兴趣的信息片段，做进一步挖掘与分析，提取出有价值的信息。而在有限指令控制系统中，系统对用户发出的特定关键词进行识别与应答，做出相应的反馈，实现指定的功能。

### 2.2.2 国内外研究现状

现在意义上的语音识别技术出现在 1952 年[20]，Davis 等人开发出了世界上的首个语音识别系统，虽然只能对少数几个单词进行识别，但依旧完成了开创性的研究。不过在二十年后，才有更多的学者投身于语音识别领域，并开展了大量的研究。这个时候的语音识别技术只能用来对样本较少的孤立词进行识别。不过这个时代的语音识别技术处于技术限制，只能对词库内的词汇进行识别，对句子以及词库以外的单词完全没有办法识别。

对于关键词识别的研究最早始于 1973 年 Bridle[21]对关键词识别的工作。不过这时 Bridle 还没有提出“关键词”这个概念。在 1977 年，Christiansen 与 Rushforth[22]两人才提出了“关键词”这个概念，并使用线性预测编码对连续语音中的关键词进行定位，取得了一些突破性的进展。这些人的工作标志着关键词识别正式从语音识别中独立了出来，成为了一个崭新的研究领域。在上世纪八十年代，Myers[23]等人将 DTW（基于动态时间规整）算法应用在了对关键词的识别之中，可以更加准确地判断关键词的位置。DTW 计算语音信号经过处理之后产生的特征向量，之后通过计算输入信号与关键词模板的特征向量之间规划距离的差值得出相似得分，根据相似得分可以在候选词中判断出关键词。不过 DTW 计算时需要过大的计算量而且无法有效利用输入的语料信息，逐渐被新的语音识别技术取代。

为了使用更加有效的方式进行关键词识别，Higgins 和 Wohlford 提出了补白模型[24]，使用模板连接的办法对语音流中的关键词进行匹配，对非关键词的信息采用另外一种方式进行建模。之后，Wilpon[25]等人将 HMM（Hidden Markov Model, 隐马尔科夫模型）引入到了关键词识别中，取得了较好的成果，对关键词识别的进展起到了极大的推动。不过随着技术的进步，一种更加优秀的模型取代了 HMM。在 2009 年，Hinton 在语音的声学建模中使用了 DNN（Deep Neural Networks, 深度神经网络）来代替 HMM[26]，获得了非常好的效果，从此研究人员开始探讨如何利用神经网络处理关键词识别问题。2015 年，R. Prabhavalkar 等人在存在背景噪声和远场条件下使用了 DNN 技术对关键词进行了识别[27]，可以改善小尺寸关键词识别中的鲁棒性，很大程度上提高了关键词识别的准确率。与之前的方法不同，DNN 不再像 HMM 一样对语音的前后转移概率进行假设，而在对相邻语音帧的拼接中记录下了语音数据的时域信息，提升了模型的分类效率。

不过为了进一步提高语音识别的准确率，需要考虑其上下文信息，建立语言模型。RNN（Recurrent Neural Network,递归神经网络）可以考虑到更多的上文信

息，具有更好的表现。RNN 通过以显式方式对时间信号依赖性建模，从嘈杂语音中预测纯净语音[28]。RNN 具有良好的非线性建模能力，可以一定程度上削弱噪声的干扰，已经证明了自己是一种非常有效的噪声清除方法。但随着网络层数的增加，RNN 容易出现梯度消失的问题，使得后面的层难以利用上文信息，导致识别准确率下降。为了解决这个情况， LSTM（Long Short-Term Memory,长短期记忆网络）被开发了出来。LSTM 作为 RNN 的一种，增加了输入门、输出门和遗忘门，可以更好的在相邻节点之间传递状态，能够有效地解决梯度消失的问题，可以更好地利用上文信息。此外， Bi-LSTM（Bi-directional Long Short-Term Memory,双向长短期记忆）进一步改善了这种结构[29]，它可以更有效地利用时间上下文信息，从而实现从嘈杂语音到纯净语音的特征映射，使得语音识别的准确率有一定增加。

与此同时，另外一种语音识别方法也在发展，那便是语音识别的端到端方法。语音识别的端到端方法核心思想是考虑输入和输出语音之间的顺序是否对应，不需要考虑时间上是否对应，致力于改变系统的损失函数来解决这一问题[30]。端到端技术主要分成两类：一类是 CTC 方法，另一类是 Sequence-to-Sequence 方法。CTC 方法由 Graves 等人提出[31]，为一种特定的损失函数，不需要预先将数据对齐，使用一个输入序列与一个输出序列就可以训练。Sequence-to-Sequence 由 Google 应用于语音识别领域。由于其性能和灵活性，Sequence-to-Sequence 模型可以应用于许多不同的情景，而无需对其原始结构进行重大修改。通过使用 Sequence-to-Sequence 的体系结构，分离的模块能够被替换为单个端到端系统。2017 年，Yanzhang He 等人将 Sequence-to-Sequence 模型与 RNN-T 模型相结合[32]，完成了一个共同学习声学和语言模型组件，可以将音素或字素作为子词单元进行预测，从而使该组件能够检测任意关键词词组，而无需任何词汇外的词。

在国内，对关键词识别的研究虽然起步较晚，不过在国内研究人员的努力下，也在逐步追赶国际前沿。在国内对关键词识别的研究中其中取得了比较大影响的有：徐明星等人提出了新的拒识方法[33]，并实现了基于音节模型的中文无限制流检测系统“Hark Man”。袁长海等人将关键词识别系统与计算机网站结合起来，制作了包含关键词识别系统的网络浏览器[34]，实现了可以使用中文关键词输入来浏览网页的方法。米尔阿迪力江·麦麦提等人使用维吾尔语对手机进行语音控制[35]，丰富了我国对少数民族语言关键词识别的研究，拓展了国内关键词识别技术应用的广度。这些国内研究对我国在关键词识别领域的空白状态有了一定的补充，一定程度上缩短了我国与国际前沿的差距，为以后的关键词识别领域的研究人员铺垫了道路。不过国内关于关键词识别系统目前依旧停留在实验室环境之中，在实际应用领域中还有较大的发展空间。

## 2.3 机器学习

### 2.3.1 机器学习模型

机器学习是一种利用现有数据的特征对未来数据的特征进行预测的方法。核心思路是将一个现实问题抽象为一个数学问题，并通过计算机对这个数学问题进行求解。机器学习一般是将已有的数据集中的特征提取出来，使用计算机程序生成的模型对这些特征进行拟合与逼近，生成一个可以用来描述这些特征的一个数学模型。在新的数据集到来时候，就可以使用已经训练好的数学模型对这个新的数据集进行预测，从而解决这个实际问题。

机器学习模型可以分为有监督学习和无监督学习两大类。这种分类的核心是在于输入的数据是否含有人工预先设置的标签。有监督学习中输入的数据需要包含标签，而在无监督学习中输入的数据不包含标签。所谓的标签，就是研究人员事先对这个数据集进行的分类，用来对关注的实际问题进行分类。在有监督学习中，模型按照标签对数据进行分类。而在无监督学习中，模型按照数据本身的特性将数据分成若干类，或者对数据的特征进行合并。机器学习在应用中根据时间顺序可以分为以下 7 步：数据收集、数据处理、模型选取、模型训练、模型评估、参数调整、实际预测。数据收集一般倾向于获取更多、覆盖范围更全面的数据。数据处理是对数据进行预处理，确定用来识别的特征，并将数据分为训练集和测试集。一个优秀的数据处理阶段可以很大程度上影响模型训练的效果。模型选择时需要根据实际情况选择一个适合的模型，能够用内在数学模型机制更贴近需要描述的问题。模型训练部分可以由机器独立完成，不需要人的参与。模型评估步骤主要是用测试集对预测结果进行评估，评价指标主要有精确率、召回率、$F_1$值等。之后调整参数可以使评价指标更好，使模型预测得更切合实际。最后就可以将机器学习模型应用在实际的预测中。

机器学习的效果不仅取决于输入的数据，选择的机器学习模型是否适合这些数据也影响着机器学习的输出结果。因此，选择一个适合于解决这个问题的模型就很重要。有监督学习模型可以分为分类和回归两大类。分类面对的是离散变量，目标是将输入数据打上标签，为一个定性过程。回归面对的是连续变量，目标是根据输入数据给出其预测的结果值，为一个定量过程。关键词识别为一个分类问题，常用的机器学习模型有线性回归模型、树形模型以及支持向量机。下面对这些机器学习模型分别进行介绍。

逻辑回归模型(Logistic Regression)为一种线性回归模型，核心思想为假设数据服从 Logistic 分布，然后对参数进行估计。Logistic 分布为一种连续型概率分布，其分布函数和密度函数分别为：

$$F(X) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}} \tag{2-3}$$

$$f(X) = F'(X \le x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-\frac{x-\mu}{\gamma}})^2} \tag{2-4}$$

逻辑回归模型的目标是寻找到分类概率$P(Y=1)$与输入向量$x$的直接关系，然后使用分类概率来进行分类。$w^T x + b$表示样本点所处的位置，可以使用$y$的类后验概率估计对这个样本点位置进行拟合，两者可写为：

$$w^T x + b = ln\frac{P(Y=1|x)}{1-P(Y=1|x)} \tag{2-5}$$

$$P(Y=1|x) = \frac{1}{1+e^{-(w^T x+b)}} \tag{2-6}$$

为了解其中的参数$w$，需要计算其最大似然函数。极大似然估计方法是为了寻找到一组参数，在这组参数下，数据的似然度最大。似然函数$L(w)$可以表示为：

$$L(w) = \prod[p(x_i)]^{y_i}\,[1-p(x_i)]^{1-y_i} \tag{2-7}$$

其损失函数$J(w)$可由下式计算：

$$J(w) = -lnL(w) = -\frac{1}{n}(\sum_{i=1}^{n}(y_i lnp(x_i) + (1-y_i)\,ln(1-p(x_i)))) \tag{2-8}$$

优化的主要目标是找到一个梯度下降的方向，当参数朝这个方向移动时能够使得损失函数的数值下降。不过这样做法可能会导致过拟合，即陷入局部最优而无法寻找到全局最优。在实际中，常常使用正则化来避免过拟合，即在损失函数后面加入一项正则化因子。加入了正则化因子之后，模型会倾向于学习较小的权重，对较大权重的学习会受到抑制，这样可以减小过拟合发生的概率。损失函数$J(w)$表示为：

$$\mathrm{J(w)} = -\frac{1}{n}(\sum_{i=1}^{n}(y_i lnp(x_i) + (1-y_i)\,ln(1-p(x_i)))) + \lambda\|w\|_2 \tag{2-9}$$

其中$\lambda$为正则化系数。对参数$w$进行求解的常用方法为随机梯度下降法。随机梯度下降法使用损失函数对参数$w$的一阶偏导来找下降方向，每次求完偏导之后就对参数$w$进行更新，更新方式如下：

$$g_i = \frac{\partial J(w)}{\partial w_i} = (p(x_i) - y_i)x_i \tag{2-10}$$

$$w_i^{k+1} = w_i^k + g_i \tag{2-11}$$

每次更新参数$w$后，比较$\|\mathrm{J(w^{k+1})} - \mathrm{J(w^k)}\|$，如果小于规定阈值，则停止迭代，输出参数$w$。

岭回归模型(ridge regression)也是一种线性回归模型，与逻辑回归模型本质的区别在损失函数的不同。岭回归模型的损失函数如下：

$$\mathrm{J(w)} = \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\|w\|_2 \tag{2-12}$$

岭回归模型通过损失部分信息的方法，减小了过拟合的情况，使得回归系数更符合实际，并且更加可靠。

决策树模型为一种树形模型，包含一个根节点与若干个内部节点和叶节点，采用迭代的方式进行构建。其各个节点的功能与内容如下：根节点：包含所有样本；内部节点：特征属性的判断；叶节点：决策的结果。决策树模型的构建分为三个步骤：特征选择、决策树生成、剪枝。特征的选择决定了决策树采用哪些特征作为输入特征来在内部节点进行判断。在决策树生成步骤，每一个样本从根节点出发，在每个内部节点进行评价指标的计算，按照要求被分到下一个节点，直到生成全部叶节点或者符合预定的要求为止。剪枝的目的是为了降低出现过拟合的风险，通过主动去掉某些相连的内部节点和叶节点来实现。在预测时，每个样本从根节点出发，经过内部节点的判断，最终被分入特定的叶节点，产生分类结果。按照对特征进行选择的方式不同可以将决策树划分为以下 3 种：ID3、C4.5 和 CART。本文中采取 CART 算法。

随机森林模型为一种基于决策树模型的改进型，有效地解决了决策树中经常出现的过拟合问题。随机森林的主要思想是构造大量的决策树进行运算，在数据输入进来后，每个决策树都进行计算。最后将分类的输出结果最多的一类当作最终的输出结果。

极端随机树模型为随机森林模型的一个变种。不同点为在对每棵决策树进行训练的时候，随机森林模型将对训练集的随机采样作为训练决策树的输入，而极端随机树模型选择训练集整体作为训练决策树的输入。因此极端随机树的应用面更加广泛，能够适用于样本偏差更大的数据中，具有更小的方差。

支持向量机(SVM)为一种二分类模型，核心思想是将实际问题抽象为一个凸二次规划问题，并对其求最优化。SVM 初始模型为一个从模式识别中发展而来的分类器，由苏联学者 Vapnik 和 Lerner 在 1963 年提出[36]。基本模型为一个线性分类模型，在采用不同的核函数之后可以进行非线性分类。

对于训练集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\dots,(\boldsymbol{x}_N,y_N)\}$，其中 $\boldsymbol{x}_i\in\mathbb{R}^n, y_i\in\{-1,+1\}, i=1,2,\dots,N$，$\boldsymbol{x}_i$为第$i$个特征输入向量，$y_i$为第$i$个标记，当它等于+1 时为正例；为-1 时为负例。对于一个超平面$\boldsymbol{w}\cdot\boldsymbol{x}+\boldsymbol{b}=0$，样本点$(x_i,y_i)$到其的几何距离为：

$$\gamma_i=y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot\boldsymbol{x_i}+\frac{b}{\|\boldsymbol{w}\|}\right) \tag{2-13}$$

该样本点到这个超平面的最近距离$\gamma=\min\limits_{i=1,2,\dots,N}\boldsymbol{y_i}$。因此 SVM 的最优化问题可以表述为：

$$\max_{\boldsymbol{w},b}\gamma \quad s.t. \quad y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot\boldsymbol{x_i}+\frac{b}{\|\boldsymbol{w}\|}\right)\geq\gamma, i=1,2,\dots,N \tag{2-14}$$

对上式的左右同时除以$\gamma$，并令$\boldsymbol{w}=\frac{\boldsymbol{w}}{\gamma\|\boldsymbol{w}\|}$，$b=\frac{b}{\gamma\|\boldsymbol{w}\|}$，上式可简化为：

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}+b)\geq 1, i=1,2,\dots,N \tag{2-15}$$

此时，最大化$\gamma$等价于下式最小化：

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^{2} \quad s.t. \quad y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}+b)\geq 1, i=1,2,\dots,N \tag{2-16}$$

为了对上式进行求解，需要构造拉格朗日函数，如下式：

$$L(\boldsymbol{w},b,\boldsymbol{a})=\frac{1}{2}\|\boldsymbol{w}\|^{2}-\sum_{i=1}^{N}a_i(y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}+b)-1) \tag{2-17}$$

为了对该拉格朗日函数进行求解，令$L(\boldsymbol{w},b,\boldsymbol{a})$对$\boldsymbol{w}$和$b$的偏导为 0，得到下式：

$$\min_{\boldsymbol{w},b}L(\boldsymbol{w},b,\boldsymbol{a})=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_i y_i y_i\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right)+\sum_{i=1}^{N}a_i \tag{2-18}$$

考虑到实际训练数据不能做到线性完全可分，支持向量机引入了损失函数，可以将原式转化为下式：

$$\begin{aligned}&\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^{2}+C\sum_{i=1}^{m}\xi_i\\ &s.t. \quad y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}+b)\geq 1-\xi_i\\ &\xi_i\geq\boldsymbol{0}, i=1,2,\dots,N\end{aligned} \tag{2-19}$$

其中$\xi$为松弛变量，$\xi_i=\max(0,1-y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}+b))$为损失函数，$C>0$为惩罚系数，$C$越大，对分类的就惩罚越大。通过对这个方程进行求解，就可以得到分类的超平面$\boldsymbol{w}^{*}\cdot\boldsymbol{x_i}+b^{*}=0$。分类决策函数$f(\boldsymbol{x})=sign(\boldsymbol{w}^{*}\cdot\boldsymbol{x_i}+b^{*})$。之后接收到新的输入后，计算其决策函数的值，和阈值进行比较之后，就可以实现对其的分类。

### 2.3.2 机器学习在关键词识别中的应用

在本世纪初，基于 HMM 内核的语音识别模型已经有了很大的进展，不过对于诸如词库外的单词，错误的开头输入等问题还很难处理。此外该模型在流利性上以及实际听觉体验上还有不少的问题[37]。为了处理这些问题，以 Benayed 为代表的一些学者开始将机器学习引入到关键词识别之中[38][39]。通过比较置信度与精确率，证明使用支持向量机作为机器学习模型用来进行关键词识别是可行的。不过在这里 Benayed 等学者还只是对实现保存好的录音文件进行识别，不能对实时的语音流进行识别。在国内，陈太波与张翠芳将支持向量机与 HMM 相结合[40]，并且引入了融合参数，有效地提高了关键词识别的精确率。

然而这些基于机器学习的关键词识别系统都是针对事先录制好的音频文件进行识别，不能对实时输入的语音流进行识别，只能适用于实验室的环境，不适合应用在实际生活工作中。

### 2.3.3 评价指标选择

对于关键词识别系统的评价，传统上有精确率(Precision)和召回率(Recall)两个评价指标。这两个评价指标从不同的角度反映了识别的准确性，精确率表示的是正样本被识别为正类的数量与所有被识别为正类的比例。召回率表示的是正样

本被识别为正类的数量与所有识别正确的样本的比例。精确率和召回率从不同的角度反映了识别的准确性，不过在实际评估应用中，精确率和召回率的变化情况通常是不同时的。一个指标的增加通常会导致另外一个指标降低。如果要综合考虑这两个指标，可以使用综合评价指标$F_1$，$F_1$同时考虑了以上两个评价指标，可以有效地描述机器学习模型的识别效果。当$F_1$较高时，说明系统有效性比较高。

对于实时语关键词识别，还有一个重要指标，就是识别速度。实时关键词识别要求在讲话人说话的同时就进行关键词识别，要求有较强的实时性，一旦识别速度无法满足要求，还会出现较大的延迟，更有甚者会导致输出的错位。因此，识别速度过慢会影响识别效果以及观看体验。同时，识别速度只要满足一定的门限，可以正好处理输入的语音就可以，识别速度过快也会导致对资源的浪费与识别准确度。

## 2.4 DMX512 设备

### 2.4.1 串口通信

串口通信为一种应用广泛的以比特位为收发信息基本单位的通信方式。在通信中，虽然串口通信这样按位传输信息比按字节传输信息的方式比较缓慢，但是串口通信不需要位同步，控制简单，易于管理，故障率低。串口通信中主要的参数有波特率、数据位、停止位和奇偶校验位。波特率用来衡量每秒传输的码元个数。数据位用来携带传输的数据。停止位用来标识数据位的结束。奇偶校验位用来减少传输中出错的概率。在两个端口进行通信前，必须要保证这两个端口之间的参数完成了匹配，这样才能正常地通信。

串口通信的基本标准为 RS-232，相互通信的两个设备之间需要采用相同的标准才能通信，用于保证不同设备之间通信稳定。不过在 DMX512 协议中串口通信使用的协议为 RS-485，相较于 RS-232 增加了联网功能。

### 2.4.2 DMX512 协议

DMX512 协议是一种数据调光协议，由美国舞台灯光协会提出。协议提出了一个供灯光控制器与灯光设备之间通信的协议标准。在实际应用中，一个 DMX 接口最多可以控制 512 个通道，一个灯光设备需要由十几个通道控制，这意味着一个 DMX 设备可以同时控制数十台灯光设备，如何确保 DMX 设备与不同的灯光设备之间稳定、准确地通信成为了 DMX512 协议产生的重要原因。DMX512

协议规定使用数据帧进行传输，一个 DMX512 数据包最多可包含 512 个数据帧。DMX512 数据包包含数据包之间的若干位空闲信号(MTBP)、一个长度为 22 位工作在低电平的中断标志信号(BREAK)、2 位高电平信号(MAB)、8 位低电平起始码(SC)、若干个 11 位数据帧，每位占用的时间为 4 μs。其中数据帧包含 1 位起始码、8 位控制信号、2 位高电平结束标志位。DMX512 协议使用串行通信的方式收发数字信号，为全双工模式。这样，只需要一根信号线便可以实现控制台和灯光设备之间信息的传递，这样便减少了灯光控制台和灯光设备之间的线路数量。同时由于 DMX512 控制协议的提出以及规范地使用，使得使用了这个协议的各个灯光设备与灯光控制台之间能够互相连接，提高了灯光设备的兼容性[41]。

DMX512 协议数据包如图 2-2 所示：

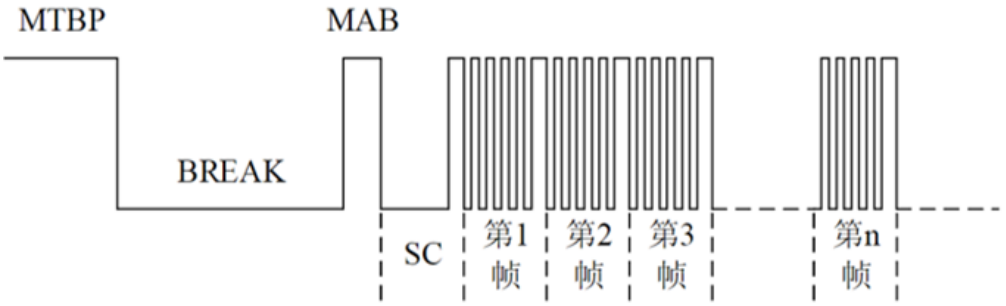


图 2-2 DMX512 协议数据包格式

Figure 2-2 DMX512 protocol data packet format

最后介绍一下 DMX512 设备。DMX512 设备用于连接计算机主机与灯光设备，使计算机主机可以通过 DMX512 设备对灯光设备发送指令，同时，灯光设备也可以经由 DMX512 设备向计算机主机传送自己的状态信息。DMX512 设备采用 RS-485 串口，异步工作，可以保证信号在计算机主机与灯光设备之间的双向传播。在进行 DMX512 时需要提前设置好读取方式、数据缓存空间、等待接收数据的时间等内容。

## 2.5 本章小结

本章介绍了本系统所需要的相关技术原理。对语音信号的预处理方法与 MFCC 特征提取的方法进行了介绍。对语音识别的历史以及关键词识别技术的发展做了介绍对机器学习模型、机器学习在关键词识别中的应用方式以及机器的评价指标进行了介绍。对串口通信原理与 DMX512 协议进行了介绍。为后续工作的进行做了技术上的铺垫与支持。

# 3 系统的总体设计和实现

本章对本智能实体沙盘系统的框架构造与实现方法进行了介绍。首先介绍了系统提出的原因。叙述了现在智能实体沙盘的不足之处，介绍了智能实体沙盘系统的实体结构以及系统即将面临的挑战。其次介绍了系统的实现方法。在这部分叙述了对系统添加功能并进行改进的思路，同时对系统的总体结构进行了介绍。之后对系统的四个模块分别进行了详细的介绍，介绍了这些模块的原理与可以为系统实现的功能。最后介绍了系统的编程实现。

## 3.1 智能实体沙盘系统

### 3.1.1 系统的提出

智能实体沙盘目前在军事领域以及诸多民用领域发挥着广泛的作用。在军事指挥中，指挥员可以通过智能实体沙盘观察战场形势，听取汇报展示，也可以调度资源，制作决策，下达指令[42]。在民用领域中智能实体沙盘最常用于教学演示领域。比如可以模拟多种天气与多种地形，也可以对自然灾害进行模拟，实现多种丰富的人机交互功能[43]。

本文所研究的智能实体沙盘系统工作在军事指挥领域，由讲解员进行讲解与展示，听众进行收听与提问。目前，讲解员在对传统的军事沙盘进行讲解时，需要中央控制台的控制人员人为对播放进行控制，同时播放的视频只能按照预定好的时间线顺序进行播放。在这种情况中，讲解员与中央控制台的控制人员难以沟通，十分不灵活，无法根据演示现场的实际情况进行变化。此外由于讲解员和中央控制台控制视频播放的人员相互独立，在进行沙盘模拟演示时会出现沟通不及时而导致播放出现错误。

针对以上需求，本智能实体沙盘系统针对以上问题进行了改进，拟为智能实体沙盘增加语音识别功能，讲解员可以直接通过语音来对智能实体沙盘的灯光设备进行控制，提高用户的体验效果，目的是在一定程度上代替中央控制台控制播放设备的人员完成相应的工作任务，从而灵活地实现播放的视频在时间顺序上的调用。

目前国内对于为智能实体沙盘添加语音识别模块或者关键词识别模块的工作直到近几年才有人进行了研究，而且研究的人数较少，很少有具有建设意义的工作。比如在 2017 年，郭永刚将离线语音识别模块嵌入到基于 STM32 的智能实

体沙盘中[44]，用来将识别到的文本信息发送到服务器中。同样在 2017 年，姚星辰等人使用关键词识别的方法实现对智能实体沙盘系统视角与单位的控制[45]。除了研究领域内，目前的一些智能实体沙盘已经开始在实际工程应用中使用到语音识别技术，能够有效地提高智能实体沙盘的交互性。例如中天智领开发出的智能实体沙盘，采用云语音识别技术，能够通过语音识别进行交互。按下激光笔上的语音按键，说出特定的关键词，智能实体沙盘就可以调用相关监控或者计算进来进行相应的操作，提高了整体的调度水平。从这些研究中可以看到，目前对于智能实体沙盘的语音控制主要有两种选择：一种是使用连续语音识别，一种是对关键词进行识别。考虑到本智能实体沙盘中的灯光设备能够进行的动作有限，只需要使用少数几种指令就可以满足其功能。针对这种需求，本系统决定采用关键词识别的方法。

但是军事沙盘的一大特点就是保密性，这点导致了军事沙盘不能与外界进行通信，也不能接入外部网络。在上述三个智能实体沙盘系统中，语音识别模块或者关键词识别模块都依赖于网络中的语音识别平台，无法离开网络独立运行。同时在目前对于应用在实际系统中的关键词识别模块，大多都采用网络上的语音识别平台，无法做到本地识别，不能满足本智能实体沙盘的基本需求。而对于在工作在本地的关键词识别模型，很少有应用在实际系统中的情形。此外，目前的前沿关键词识别技术基本使用的都是基于神经网络技术的方法，需要强力的硬件设备作为支持，本智能实体沙盘系统由于是一个轻量型的系统，难以使用强力的硬件来支持复杂的技术。同时，本智能实体沙盘应用的场景为实时的关键词识别，也不是传统上的离线关键词识别，对于关键词的识别速度有很大要求。因此就需要一种可以工作在本地、规模小、识别速度快的工具来进行关键词识别的工作。

可以看到，目前现有的智能实体沙盘无法满足本系统的实际需求。为了丰富目前场景中智能实体沙盘的功能，本文提出了一个实时关键词识别的智能实体沙盘灯光控制原型系统，可以实现上文所需求的功能。

### 3.1.2 智能实体沙盘的结构

本智能实体沙盘系统原本是一个集图像采集处理、POI 域选择、光斑识别与检测等功能于一体的激光光斑识别系统。功能是利用摄像机追捕激光笔实时发出的光斑，并将光信号转化为数字信号传递给计算机主机。计算机主机接收到这个信号并计算出光斑的坐标信息，并传递给 DMX512 设备并对其进行一系列实时控制，包括指向跟踪、亮度控制、预定义的图案显示等。DMX512 设备控制灯光设备，使其照亮光斑区域，为激光笔的标记做出突出化显示，从而实现使用激光

控制智能实体沙盘灯光设备的功能[46]。

本文为其增加了对关键词识别并根据识别结果对灯光设备进行控制的功能。为了实现这一功能，本智能实体沙盘包括以下几个设备：话筒、计算机主机、DMX512 设备以及灯光设备。其中 DMX512 设备的选型为 FQSD512-PR（512 通道），该设备支持 USB 联机以及 RS232/485 转 DMX512 输出功能。DMX512 设备与计算机主机之间采用 USB 进行通信，与灯光设备之间采用 RS485 协议进行通信。灯光设备具有一组灯泡，包含 7 颗 10W 红绿蓝黄四色合一大功率 LED，能够实现 0~100%调光，水平转动角度为 540°，垂直转动角度为 180°，包含通道数量为 8/13 通道，使用 DMX512 协议进行控制。灯光设备的说明书如图 3-1 所示：

**1. 技术参数:**

输入电压: AC90~260V 50/60Hz
功耗: 100W
灯珠规格: 7 颗 10w RGBW 4 合 1 大功率 LED
灯珠寿命: 50,000 hours
调光效果: 0~100%调光
频闪: 1~20 Hz
水平/垂直转动角度: 540° / 180°
控制方式: DMX512 / 自走 /主从 /声控
通道数量: 8/13 通道
工作温度: 45°C
防水等级 IP20
包装重量: 4.5 KGS

**2. 通道说明:**

**2.1. 8 通道模式**

| 通道 | DMX 值 | 功能描述 |
|---|---|---|
| 1 | 0-255 | 水平转动 |
| 2 | 0-255 | 垂直转动 |
| 3 | 0 - 7 | 无效果 |
| | 8-134 | 调光 |
| | 135-239 | 频闪由慢到快 |
| | 240-255 | 开光 |
| 4 | 0-255 | 红色调光 |
| 5 | 0-255 | 绿色调光 |
| 6 | 0-255 | 蓝色调光 |
| 7 | 0-255 | 白色调光 |
| 8 | 0-255 | 水平/垂直转动速度（快至慢） |

**2.2. 13 通道模式**

| 通道 | DMX 值 | 功能描述 |
|---|---|---|
| 1 | 0-255 | 水平转动 |
| 2 | 0-255 | 水平微调 |
| 3 | 0-255 | 垂直转动 |
| 4 | 0-255 | 垂直微调 |
| 5 | 0-255 | 速度由快到慢 |
| 6 | 0 - 7 | 无效果 |
| | 8-134 | 调光 |
| | 135-239 | 频闪由慢到快 |

| 通道 | DMX 值 | 功能描述 |
|---|---|---|
| 6 | 240-255 | 开光 |
| 7 | 0-255 | 红色调光 |
| 8 | 0-255 | 绿色调光 |
| 9 | 0-255 | 蓝色调光 |
| 10 | 0-255 | 白色调光 |
| 11 | 0-7 | 无效果 |
| | 8-21 | 白色 |
| | 22-34 | 红色 |
| | 35-49 | 深绿色 |
| | 50-63 | 深蓝色 |
| | 64-77 | 浅蓝色 |
| | 78-91 | 洋红色 |
| | 92-105 | 黄色 |
| | 106-119 | 紫色 |
| | 120-133 | 桔色 |
| | 134-147 | 浅绿色 |
| | 148-161 | 粉色 |
| | 162-175 | 棕色 |
| | 176-189 | 金争 |
| | 190-203 | 深红色 |
| | 204-217 | 紫罗兰色 |
| | 218-231 | 深紫色 |
| | 232-255 | 16 色轮流颜色跳变 |
| 12 | 0-255 | 颜色跳变速度（慢至快） |
| 13 | 0-7 | 无效果 |
| | 8-22 | 自走模式 1 |
| | 23-37 | 自走模式 2 |
| | 38-53 | 自走模式 3 |
| | 54-67 | 自走模式 4 |
| | 68-82 | 自走模式 5 |
| | 83-97 | 自走模式 6 |
| | 98-112 | 自走模式 7 |
| | 113-127 | 自走模式 8 |
| | 128-142 | 声控自走模式 1 |
| | 143-157 | 声控自走模式 2 |
| | 158-172 | 声控自走模式 3 |

图 3-1 灯光设备说明书

Figure 3-1 Lighting equipment manual

最后完成的智能实体沙盘系统由以下部分组成：实体沙盘、计算机主机、DMX512 设备、灯光设备、工业相机、激光光标器与话筒。实体沙盘作为演示的载体，用于展示区域地形，作为激光光标器与灯光设备照明图案的受体。计算机主机作为处理与计算的核心，负责接受相机传来的光斑信息以及话筒传来的语音信号，之后对灯光投影区域坐标计算以及对关键词信息的识别，并且将控制指令通过串口通信方式传输给 DMX512 设备。DMX512 设备用于控制灯光设备，接

受计算机主机发送的指令信号，并转化为灯光的通道信息，传递给灯光设备。灯光设备则负责进行灯光照明，对激光光斑进行醒目标识，或者根据关键词做出指定的动作。工业相机负责在沙盘上获取光斑信息，并将获取的图像传递给计算机主机。激光光标器负责在沙盘上进行标注。话筒负责接受讲解员的语音信号，并将数字语音信号传递给计算机主机。本文涉及到的部分有：计算机主机、DMX512设备、灯光设备以及话筒，在图 3-2 中由红色标出。沙盘系统总体的示意图如图 3-2 所示：

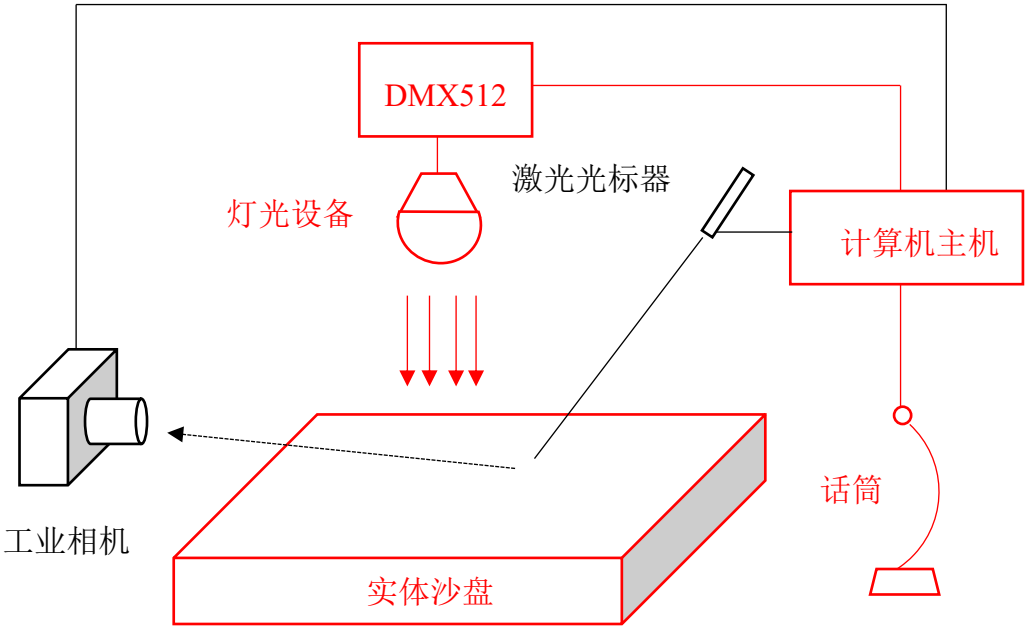


图 3-2 沙盘系统的物理结构示意图

Figure 3-2 Schematic diagram of the physical structure of the sand table system

### 3.1.3 面临的挑战

对于上文提出的智能实体沙盘系统，为了实现系统的功能，达到预期目的，更好地提高系统的性能，例如增强系统的演示效果与提高听众的观看体验，目前有如下挑战：

(1)如何在本地对关键词进行识别。目前广泛应用的关键词识别模块基本都使用第三方 API 对语音信号进行识别，因为这样比较方便省力，识别速度快，识别效果好，并且可以提高开发速度。此外本地关键词识别模块基本都是对离线的语音数据进行识别，不需要考虑实时性，虽然识别比较准确，但是识别速度慢，不适用于本场景。

(2)如何提高关键词识别的准确度。关键词识别的基本目标就是识别出说话人所讲的关键词信息，并将其转化为控制指令。目前的本地关键词识别系统受制于

资源与体积，无法做到像语音识别 API 平台一样的准确。同时，本系统需要平衡好识别速度与准确度，如果其中一个的表现不够好，就会严重影响系统的实际应用效果与用户的体验。

(3)如何提高关键词识别的速度。关键词识别另外一个评价指标就是识别的速度。由于本研究的是实时关键词识别，所以识别速度就更为重要。灯光需要快速反应，在目标区域进行动作，出现较大延迟会影响演示效果。灯光演示效果一个重要判断因素就是延迟的大小。一旦出现延迟，就会导致实际关键词与灯光动作发生错位，影响演示效果。更有甚者，由于这种错位通常是由于识别速度过慢导致的，还会对后面所有结果造成影响。因此，识别速度对于实时关键词识别是非常重要的。

(4)如何减小从识别出关键词到灯光设备做出灯光动作之间的延迟。即使关键词在正确的时间被识别出来，但是由于用户实际看到的是灯光动作，所以如果关键词与灯光动作之间有过大的延迟，将会降低用户的观看体验。

## 3.2 系统的实现方法

### 3.2.1 基本思路

本系统的特色是将本地实时关键词识别模块与智能实体沙盘结合起来，同时使用机器学习模型对关键词进行识别，对如今的智能实体沙盘系统做出了补充，能够解决上节所描述的挑战。系统设计的流程如下：

首先预先设置好要识别的关键词，从百度语音合成 API 中获取关键词对应的音频文件以及用来测试的音频文件。对关键词音频文件进行标注与编号，将测试的音频文件保存在另外一台设备中并标注出现关键词的位置与种类。

其次将这些音频文件进行预处理，之后提取出 MFCC 特征，将特征输入到机器学习模型中，对模型进行训练。比较不同的机器学习模型，从中选择出最适合本场景的模型。

之后输入包含关键词的测试用实时语音信号，经过对实时语音的处理，提取出 MFCC 特征。MFCC 特征经由机器学习模型进行计算，之后比较连续若干个关键词对应的概率判断得到输出的关键词，将其转化为指令信号，并通过串口通信传递给 DMX512 设备。

最后 DMX512 设备接收到的指令信号，根据指令信号读取灯光动作库，产生出为对应灯光的通道信息，传递给灯光设备，使其做出规定的动作。

系统基本思路如图 3-3 所示：

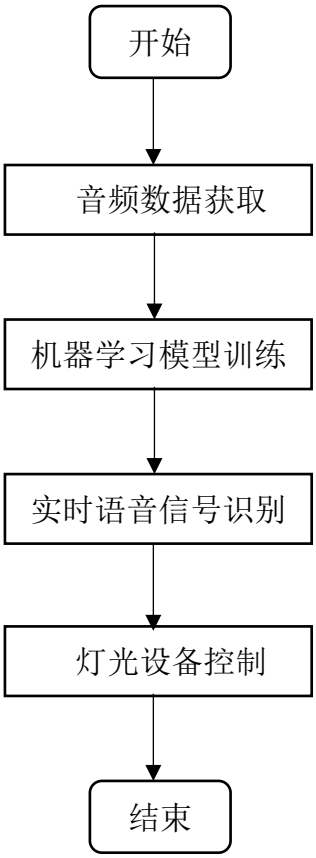


图 3-3 系统基本思路

Figure 3-3 Basic idea of the system

### 3.2.2 系统结构

根据智能实体沙盘灯光控制系统的实现流程和具体需要实现的功能，本系统可以分为训练和应用两部分。训练部分用于预先训练机器学习模型，可以划分为两个模块：音频数据获取模块与机器学习模型训练模块。应用部分用于实际的实时关键词识别，可以划分为两个模块：实时关键词识别模块与灯光设备控制模块。下面是每个模块的主要功能与实现方法。

第一部分为音频数据获取模块。该模块用于获取实验所需的关键词对应的音频文件以及在测试时需要使用实时语音文件。该模块通过调用百度语音合成 API 获得上述的音频文件。

第二部分为机器学习模型训练模块。该模块用于使用音频文件的 MFCC 特征对机器学习模型进行训练。首先对音频文件进行预处理。之后提取其 MFCC 特征作为机器学习模型训练的输入。

第三部分为实时关键词识别模块。该模块用于对实时连续语音信号中的关键

词进行实时识别。首先采用话筒作为语音信号输入端口，实时采集讲解员的语音信息，使其转换为数字信号传输给计算机主机。然后计算机主机作为控制决策设备，对实时语音信号进行语音分片、能量检测、端点检测以及预处理等操作，之后提取其 MFCC 特征并将其输入到已经训练好的机器学习模型中进行识别。机器学习模型根据概率进行阈值判断，输出识别结果中超过阈值且概率最大的一个关键词结果，并将结果转化为指令信号通过串口通信传递给灯光设备控制模块。

第四部分为灯光设备控制模块。该模块用于控制灯光设备。该模块接收实时关键词识别模块传递来的指令信号，并把指令信号转化为灯光设备的通道信息，从而控制灯光设备做出对应的动作。

本系统的总体结构如图 3-4 所示：

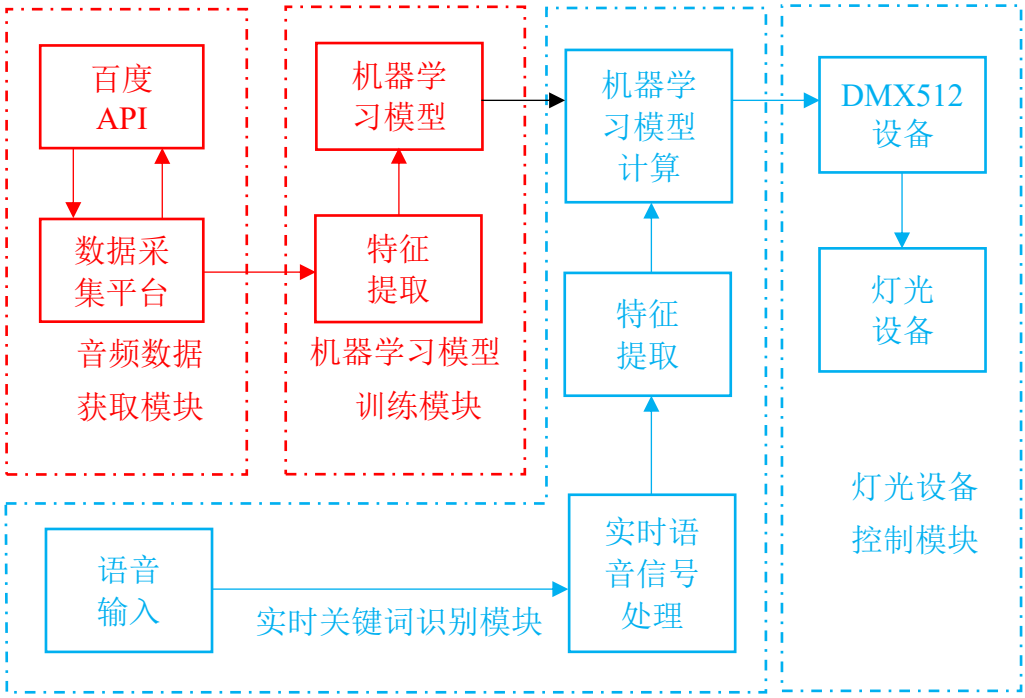


图 3-4 灯光控制系统结构

Figure 3-4 Structure of the light control system

图 3-4 中训练部分由红色标出，应用部分由蓝色标出，两者之间的联系通过训练部分将训练好的机器学习模型传递给应用部分而建立。

在宏观上，实时关键词识别模块与灯光设备控制模块的各个组成部分都是同时运行的。在微观上，实时语音的每个语音分片都需要完整地在实时关键词识别模块中经历一遍流程，如果被识别为关键词，还需要被转化为控制信号进入灯光设备控制模块。因此保证整个系统的实时性非常重要。在这些模块中，按时间顺序进入的语音信号必须按顺序地进入机器学习模型中，按系统标记的时间进入灯光设备控制模块中。系统的实际结构由图 3-5 所示：

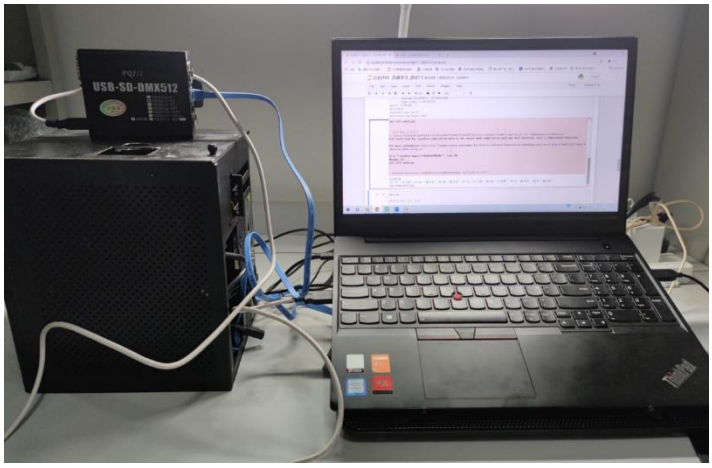


图 3-5 系统的实际结构图

Figure 3-5 The photo of the system structure

图 3-5 中计算机主机与 DMX512 设备之间通过白色数据线进行串口通信，DMX512 设备与灯光设备之间通过蓝色数据线相连接。其中 DMX512 设备不需要电源线，由计算机主机的 USB 接口供电。

本系统的核心部分在于机器学习模型的构建与应用。传统上，基于机器学习的语音识别模型主要是应用在离线的情感识别中，在实时在线的关键词识别中的应用较少。常用的对关键词识别的方法有基于神经网络，基于端到端的方法等。但是这些方法对资源开销要求大，无法做到高的实时性，不适用于本系统。由于机器学习的方法识别效率高，资源开销低，因此考虑利用机器学习的方法来进行关键词识别。本系统采用的关键词判决基于识别概率。在机器学习模型对输入的语音信号进行识别后，会产生识别为每个关键词的概率。由于对连续语音信号进行分片时，每个分片长度固定，步长很小，可以认为连续的若干个语音片段的识别结果应该是稳定的。系统对若干个连续的识别概率结果进行对应相加，如果超过设定的阈值，那么认为是一个关键词识别结果，将其中最大的一个概率值对应的关键词作为识别结果输出。如果没有超过这个阈值，就认为不是关键词部分，不输出结果，之后对下一个分片进行识别。

## 3.3 音频数据获取

### 3.3.1 关键词选择

在进行音频数据的获取之前，首先要确定识别中所用的关键词。由于系统的目的是通过识别关键词来控制灯光设备，意味着每个关键词必须对应一个独特的、有意义的灯光动作。同时，每个关键词所对应的灯光动作描述的并不是关键词本

身，而是以关键词为代表的一句话。讲解员受过训练，会在实际情况下进行配合。因此将关键词设置为每个灯光动作情景的句首词，这样可以降低从识别到关键词到灯光设备做出灯光动作的时延。只要灯光动作在这一句被该灯光动作描述的场景中做出就可以，这样就可以减小时延的问题。经过对灯光设备中灯光可以执行的动作进行筛选，本文选择了四个关键词，分别是："红方朝"、"红方向"、"双方在"以及"蓝方攻"，各对应一个灯光动作。

### 3.3.2 语音合成平台

为了获得用于机器学习训练的训练集以及在测试时使用的实时语音文件，本系统使用百度语音合成平台进行关键词数据的获取。传统的关键词识别方法在对测试关键词语音获取中，多是采集志愿者发出的声音作为数据集。本文之所以没有采用该方法原因有三：一是志愿者发出的声音灵活性差，无法对其语速或音调进行调节；二是在进行数据采集之后如果发现数据中存在问题难以更换或者补充；三是本文没有人力去获取众多志愿者的录音文件。而在百度语音合成平台可以低成本地获取大量含有关键词的音频文件，为本文的关键词音频文件获取提供了便利的条件。

百度语音合成平台提供基于 REST API 接口，适用于可发起网络请求的设备，能够将输入的文本转换为可以播放的音频文件。平台提供了 4 位发音人供选择，可以调节所需的音量、语速以及语调。本模块随机为每个关键词获得不同发音人、音量、语速以及语调若干个关键词音频文件，采用循环的方法调用 API 接口，直到获取了所有音频文件。将这些音频文件作为语音数据集保存在本地。在后续的机器学习部分中，发现训练数据集达到一定规模后训练效果不再增加，因此语音数据集的大小不需要很大，本系统使用的语音数据集包含 2400 条语音文件与 600 条非关键词的噪声文件。

在对系统的关键词识别效果进行测试评估部分，也使用了百度语音合成平台获取了 40 段长度各为 100 秒的音频文件，每个音频文件在连续语音中包含 4 种关键词，每种 2 个，总计 320 个关键词。通过使用外部设备对这些音频文件播放的方式可以用来模拟讲解员实际的讲解过程，对后面的实时关键词识别模块进行仿真与评估。

## 3.4 机器学习模型训练

这个模块用于使用音频文件的 MFCC 特征对机器学习模型进行训练。分为

三步：对音频文件进行预处理，对预处理过音频文件进行 MFCC 特征提取，将 MFCC 特征作为训练集利用机器学习模型进行训练。机器学习训练模块流程图如图 3-6 所示：

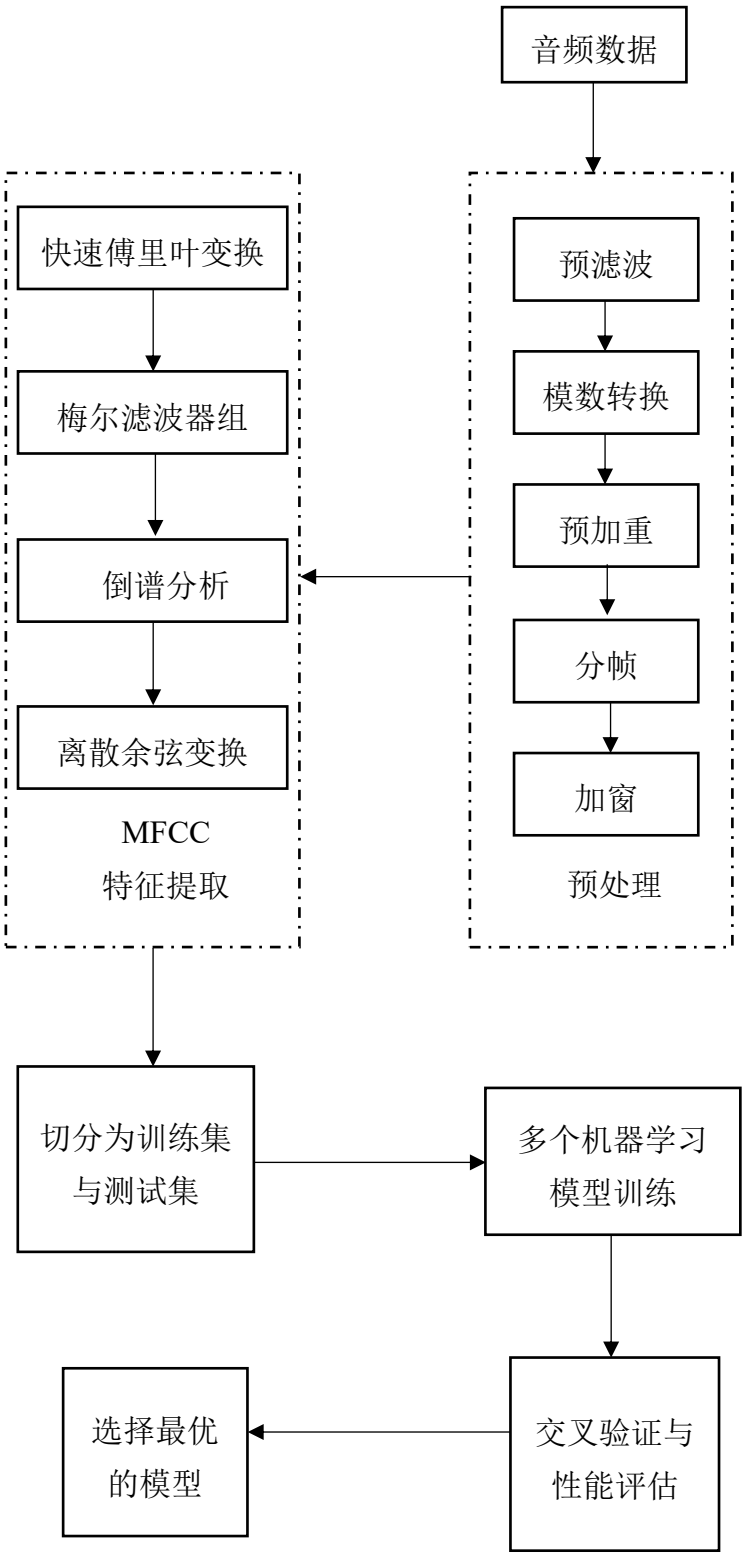


图 3-6 机器学习训练模块流程图

Figure 3-6 Flow chart of machine learning training module

### 3.4.1 预处理

对语音信号进行 MFCC 特征提取之前需要进行预处理。因为原始语音信号中包含很多对后续计算不利的干扰，需要首先进行预处理才能使得后面的语音提取和识别更加有效。预处理包含采样、预滤波、预加重、分帧、加窗五步。

采样用于获取连续的语音信号，将连续语音信号转化成计算机主机比较容易处理的数字信号。预滤波用于保留有用的语音频率部分。由于实际人说话的频率是在一定范围内的，在这个频率之外的声音一定是与系统无关的信息，需要在最开始过滤掉。预加重目的是为了增加语音的高频分辨率。因为人发出的语音信号由于口唇辐射的影响会在高频部分衰减，所以要在对其高频部分加以补偿。这里的高频指的是在人发音信号内的相对高频。分帧是为了保证后续进行特征提取部分时输入的信号在局部平稳。因为输入的信号本身是不平稳的，不过可以在一个小时隙里认为是平稳的，这小时隙大约为 10-30ms，在这里语音信号可以被看作一个准稳态过程，即语音信号具有短时平稳性。因此为了获得短时平稳的信号，我们需要将其进行分帧处理。本文中帧长取 25ms，相邻帧的重叠时间取 10ms。加窗的目的是使连续的语音信号更加连续，消除抽样对波形带来的变化，减小分帧对帧边缘处信号的损耗。本文采用汉明窗作为窗函数。

### 3.4.2 MFCC 特征提取

经过以上步骤的预处理，就可以进行 MFCC 特征的提取。MFCC 特征提取包含快速傅里叶变换、梅尔滤波器组、倒谱分析以及离散余弦变换四步。

快速傅里叶变换用来将时域上的语音信号转化为频域上，为了接下来的频域计算。FFT 作为一种适合于计算机的 DFT 计算方法，在抽样点增加时，可以大量减少计算量，有效提高计算速度。本文中 FFT 大小取 512。梅尔滤波器组为一组包含 M 个三角形的滤波器组，是根据人耳的听觉特征设计出来的，用于将实际频率转化为梅尔频率。人耳在低频段的敏感性比在高频段的敏感性强，所以梅尔滤波器组在低频部分设计得更密，在高频部分比较稀疏。这样可以减少频谱的损耗并且计算出梅尔频率。本文中 M 取 26。倒谱分析的目的是为了获得频谱的低频部分包含的包络信息。倒谱分析用于对时域信号进行傅里叶变换，再取对数，从而提取出其频谱的包络信息。包络信息分布在低频部分，细节信息分布在高频部分。我们只对其低频包络信息感兴趣，因此可以使用低通滤波器去除掉高频细节部分，只保留我们感兴趣的低频包络信息。这里音色部分产生出的对数能量就

作为 MFCC 特征的第 1 维。离散余弦变换则用来分离加性的基音信息与声道信息。最后产生出 12 维 MFCC 系数，加上对数能量，总共是 13 维 MFCC 系数。这 13 维 MFCC 系数反映了语音信号的静态特性，作为机器学习的输入特征。关键词“红方向”的 MFCC 特征频谱图如图 3-7 所示：

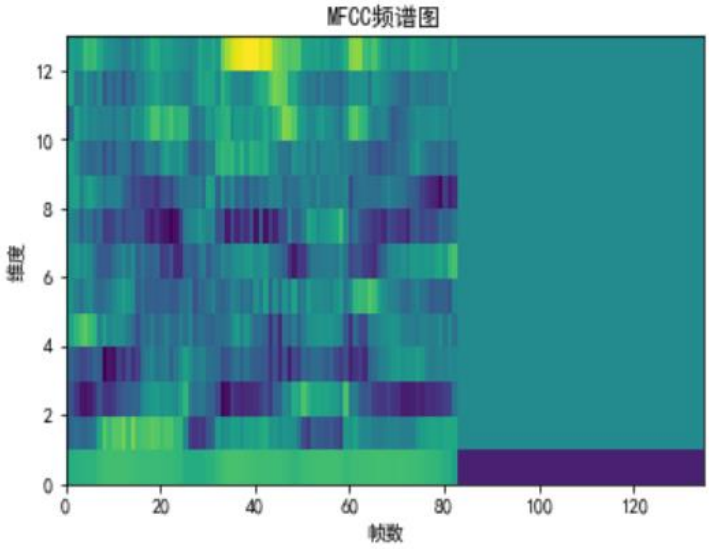


图 3-7 MFCC 特征频谱图

Figure 3-7 MFCC characteristic spectrum

图中纵轴表示的是不同维度，横轴表示的是时域中的不同帧，颜色越深表示数值越大。在后续的机器学习模型中，模型按照时间顺序，将不同帧的不同维数中的数值作为输入特征，从而完成对模型的训练。

### 3.4.3 机器学习模型的选择

机器学习模型的优劣是关键词识别的核心，常规中衡量机器学习优劣的指标主要有精确率、均方误差(MSE)等。不过本模型应用的场景为一个对实时关键词进行识别的情况，对实时性的要求很严格。一旦识别速度达不到实际的要求，会大幅度降低用户的体验，甚至使后续灯光控制步骤失去意义，会严重影响系统的使用体验。因此识别速度也就成为了衡量本系统中机器学习模型中一个重要的指标。下面根据以上评价标准选择应用在本系统中的机器学习模型。本节对逻辑回归模型、岭回归模型、决策树模型、随机森林模型、极端随机树模型以及支持向量机模型分别在识别速度，识别精确率，MSE 三个方面进行了评估。并从中选择最适合本系统的机器学习模型。

首先使用实时性来对模型进行选择。在这里使用了不同的训练好的机器学习模型对音频文件进行了识别，识别时间如表 3-1 所示：

表 3-1 不同模型的识别时间

Table 3-1 Recognition time of different models

| | 400 | 1200 | 2400 |
|---|---|---|---|
| 逻辑回归 | 8.61 | 24.73 | 47.00 |
| 岭回归 | 8.70 | 23.96 | 46.06 |
| 决策树 | 7.79 | 23.86 | 55.75 |
| 随机森林 | 9.30 | 29.62 | 59.71 |
| 极端随机树 | 9.23 | 30.03 | 66.56 |
| 支持向量机 | 27.03 | 73.19 | 158.76 |

表格第一行为语音数据集中语音文件的个数，时间单位为秒。从表中可以看到，线性模型与树形模型识别时间相差不大，这是由于在这些模型中系统的大部分计算开销都花费在了对 MFCC 特征的提取上所导致的。同时，线性模型与树形模型都可以满足系统对于识别速度的要求，这就为对模型的优化提供了富裕度。因为只要识别速度能够满足最低门限就可以符合系统标准要求，在此基础上可以对机器学习模型进行性能上的优化，从而有更好的识别效果，提升实际识别中的精确率。但是支持向量机的识别时间远远超过线性模型以及树形模型，这是由于算法本身的复杂性导致的。同时支持向量机的识别时间都超过了系统要求的最大识别时间门限，表明了支持向量机是无法在实际系统中应用的。即使后面支持向量机的精确率或者 MSE 再优秀也不能使用，因为它无法满足系统的实时性需求。

在实际关键词识别中，由于语音分片间隔为 0.04 秒，意味着 1 秒内机器学习模型需要识别 25 个音频文件，这是最低门限时间。实际识别的语音分片长度与训练用的音频文件时间长度相同，都为 2 秒。而且实际关键词识别时还存在着其他会导致实时性降低的情况，比如从关键词说完到识别出关键词所产生的时延，从识别出关键词到传递给灯光设备所产生的时延，灯光设备接收到指令信号到灯光设备做出动作所产生的时延等等。此外还存在着其余的资源调用的情况，比如实时语音信号的获取与灯光设备的调用都会占用系统的资源，无法满足在实际识别过程中关键词识别部分能被分配到理想的计算资源，会影响关键词识别效果。因此希望识别时间能够在保证识别精确率大致不变的情况下尽量缩短，可以降低时延，减小资源占用与开销。这也是实时系统中的基本需求。

精确率是衡量机器学习模型效果的重要参数，用来衡量机器学习模型识别结果与真实结果的相似性。下面将原始音频数据中的四分之三作为训练集，四分之一作为测试集。使用训练集对机器学习模型进行训练，使用测试集对所得到的结果进行测试，计算不同模型的精确率，结果如表 3-2 所示：

表 3-2 不同模型的精确率

Table 3-2 Accuracy of different models

| | 400 | 1200 | 2400 |
|---|---|---|---|
| 逻辑回归 | 0.99 | 0.93 | 0.99 |
| 岭回归 | 0.98 | 0.93 | 0.99 |
| 决策树 | 0.98 | 0.91 | 0.92 |
| 随机森林 | 0.97 | 0.88 | 0.91 |
| 极端随机树 | 0.96 | 0.92 | 0.91 |
| 支持向量机 | 0.98 | 0.90 | 0.95 |

表格第一行为语音数据集中语音文件的个数。可以看到，线性模型在测试集大小变化时都是最优的，支持向量机其次，树形模型精确率最低。不过在模型实际应用过程中，输入的实时语音信号无法被训练集的数据全部代表，如果精确率过高，可能是发生了过拟合的情况。而树形模型的精确率处在 0.9 左右，是由于树形模型采用了剪枝的方法一定程度上规避了过拟合的发生，并不是说明树形模型因为精确率过低而不能在系统中应用。

对 MSE 的评估通常使用交叉验证来实现。交叉验证是一种有效对不同机器学习模型进行评价的方法，能够有效评估模型的识别性能，被广泛应用在对机器学习模型性能的评估中。在对原始训练集进行训练时，一般会将原始训练集拆成一个训练集和一个测试集，通过用训练集的数据来训练机器学习模型，然后用测试集的数据来对模型进行评估，这种方式可以保证训练的效果。因为在进行机器学习模型训练时，训练得到的模型可能只会对训练集有正常的识别效果，但是对于外部实际使用的数据集工作地很不理想。这是因为机器学习模型对于初始的条件很敏感。因此需要把原始训练集的一部分单独拿出来作为测试集来客观测试机器学习模型的拟合程度。

交叉验证的基本思想是在机器学习模型训练时对数据集多次划分，从而选取最佳模型。最常用的交叉验证方法为 K 折交叉验证，其得名的原因就是其将整个数据集分为 K 份来进行交叉验证。其步骤如下：首先将整个数据集均分为 K 份，选取 K-1 份作为训练集，剩下的 1 份作为测试集，这样训练一个模型，并计算其 MSE（均方误差）。之后不停地重复这个过程，直到每一份都被当作过测试集。最后计算这些 MSE 的平均值，来评估这个模型的特性。

交叉验证是为了验证不同模型的适用程度而产生的，并不能提高指定模型的效果，但是能在比较的不同机器学习模型中选择泛用性最高的一个。因为在每次在对训练集进行训练与计算时，并不会因为交叉验证而优化训练模型的参数，最

后生成的模型也不是 MSE 值最小的模型。在下面，利用交叉验证对逻辑回归模型、岭回归模型、决策树模型、随机森林模型、极端随机树模型、支持向量机模型进行了评估，结果如表 3-3 所示：

表 3-3 不同模型的 MSE

Table 3-3 MSE of different models

| | 400 | 1200 | 2400 |
|---|---|---|---|
| 逻辑回归 | 0.01 | 0.19 | 0.03 |
| 岭回归 | 0.02 | 0.18 | 0.03 |
| 决策树 | 0.02 | 0.28 | 0.33 |
| 随机森林 | 0.03 | 0.46 | 0.44 |
| 极端随机树 | 0.02 | 0.27 | 0.23 |
| 支持向量机 | 0.02 | 0.26 | 0.15 |

表格第一行为语音数据集中语音文件的个数。可以看到，如逻辑回归、岭回归这样的线性模型表现最好，支持向量机表现次之，决策树、随机森林、极端随机树这样的树形模型较差。即线性模型的泛用能力比较强，可以灵活地在不同情景下运行。

综上所述，本系统决定采用逻辑回归模型与决策树模型作为机器学习模型，并在第四章比较这两个模型在实时关键词识别时的性能。

使用 Python 语言程序对该模块进行编程，该模块伪代码如表 3-4 所示：

表 3-4 机器模型训练算法

Table 3-4 Machine model training algorithm

**算法 1. 机器模型训练**

**输入：音频文件**

**输出：机器学习模型**

1. for i from 0 to 400 do
2. 从第 i 个音频文件中提取语音信号$sig_i$
3. 将$sig_i$中小于阈值的元素置零
4. 在$sig_i$列表右侧增加 0 元素，补充到规定长度
5. 对$sig_i$进行预处理
6. 从$sig_i$中提取$mfcc_i$
7. end for

表 3-4 机器模型训练算法（续）

Table 3-4 Machine model training algorithm(extend)

| |
|---|
| 8. $mfcc[] \leftarrow mfcc_1, mfcc_2, \dots, mfcc_{400}$ |
| 9. 使用$mfcc[]$训练机器学习模型 |
| 10. 导出训练好的机器学习模型 |

## 3.5 实时关键词识别

该模块用于对实时语音信号进行识别。分为三步：对实时语音信号进行处理，使用训练好的机器学习模型进行识别，指令转化。实时关键词识别模块流程图如图 3-8 所示。

### 3.5.1 实时语音信号的处理

对实时语音信号的处理方法与对关键词音频文件的处理办法有相同之处与不同之处。相同之处在于它们都需要进行 MFCC 特征的提取；不同之处在于对实时语音信号的特征的提取前需要额外进行语音分片、时间标记、能量检测以及端点检测工作。

语音分片是指对话筒接收到的实时语音信号进行切割保存的工作。如果不将实时信号保存到本地，机器学习模型就无法对其进行处理。同时切片的时间宽度必须不能超过关键词音频文件中最宽的那一个，这样才能不遗漏切片中的信息。为了与机器学习模型进行对齐，需要对每个切片后面补零。同时，如同加窗一样，对实时语音信号进行切片也会遗漏其两端的信息，很可能使一个关键词被分为两半，导致无法识别出结果。因此在进行语音分片时，相邻的两个语音片需要有一定的重合，这样可以增加关键词识别的效果。本系统中一个语音分片的长度为 2s，步长为 0.04s，相邻两个语音分片之间有 1.96s 的重叠。为了保证录音最开始产生的语音分片的长度也是 2s，系统在第 3s 的开始输出并保存这些语音分片，每秒产生 25 个语音分片。

时间标记是为了保证实时性而进行的工作。在进行语音分片后，系统将这些语音片段保存在本地，并按顺序进行编号。由于语音分片的时间间隔是固定的，利用编号就可以确定该语音分片的输入时间，从而与输入语音作对比，评估关键词识别的精确率。也可以在机器学习输出结果前，判断时域内临近的语音分片，用来进行识别概率的相加。

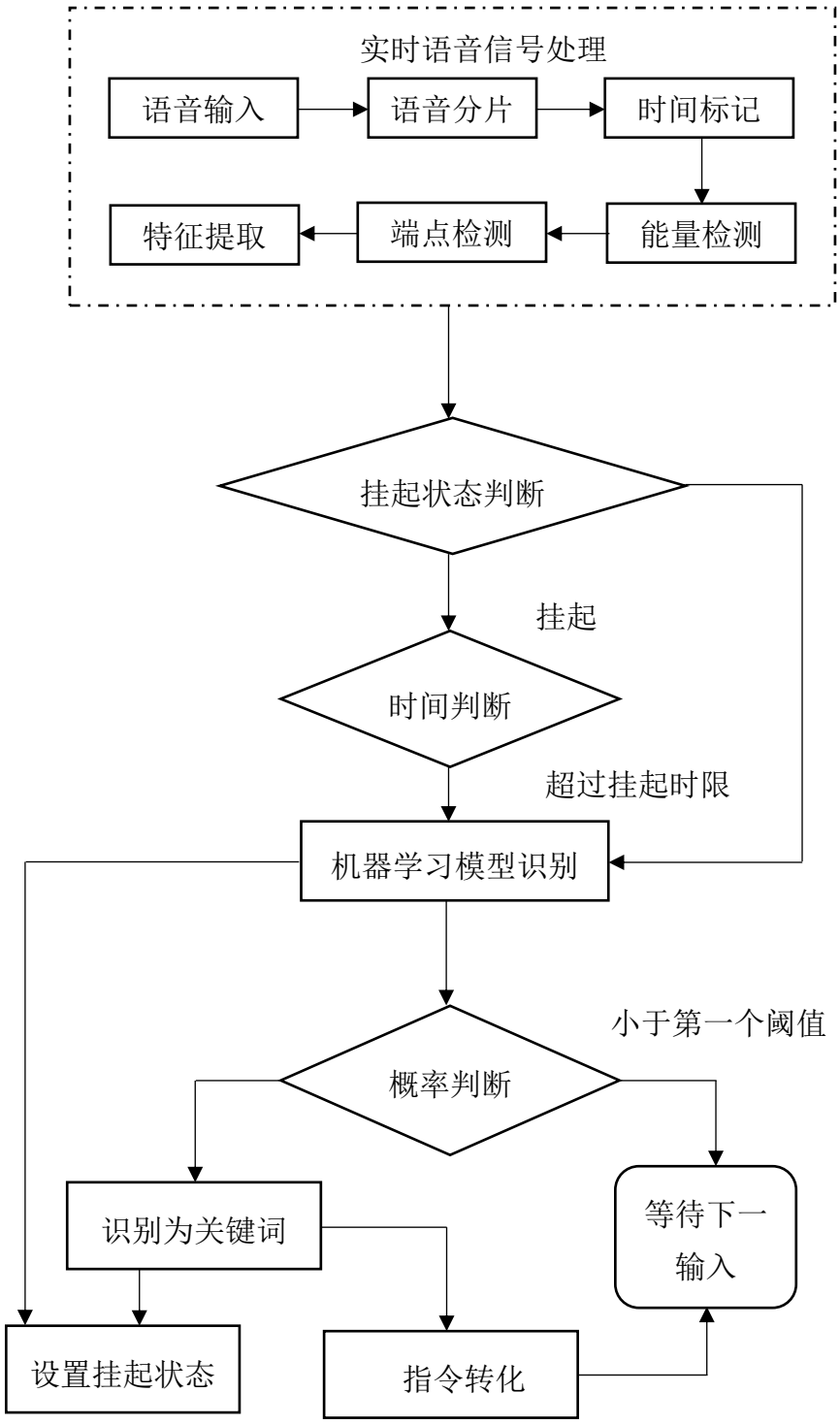


图 3-8 实时关键词识别模块流程图

Figure 3-8 Flow chart of real-time keyword spotting module

能量检测部分用来过滤掉连续静音部分。语音信号的短时能量为该语音区间内样本值的加权平方和，可以作为区分有声和无声的依据。对于原始的关键词音频文件，这些语音信号的短时能量都超过了一定的门限。可以认为，如果同样长度的语音分片的短时能量达不到该门限阈值，那么可以认为这个输入语音信号片

段不是包含讲解员说话的信号，可能是讲话中的停顿或者噪声信号。在这一步先将其滤除，可以减少识别负担，提高识别效率。同时，对于通过该门限的语音分片，还要进行一个对样本值的过滤，用来过滤语音分片中的噪声。这里对样本值设置一个门限阈值，如果达不到，则认为是环境噪声，系统将这个样本值置零；如果能达到，则认为是语音信号，在这步不再进行处理。

端点检测的作用是确定一段连续语音信号的起始部分和结束部分，这样就可以将这个连续语音信号分成两类，一类是语音部分，一类是非语音段部分。因为只有语音部分是我们感兴趣的部分，对于非语音段部分没有识别的必要。在能量检测部分，系统已经将达不到阈值的样本值置为零。在这步，系统选择语音分片的第一个非零样本值点，将其作为起始点；系统选择语音分片的最后一个非零样本值点，将其作为终止点。

进行了如上操作之后，对语音分片信号进行预处理与 MFCC 特征提取，方法与机器学习模型训练模块使用的方法相同。

### 3.5.2 关键词识别

关键词识别部分为本模块的核心，也是影响整个系统的关键。关键词识别部分使用已经训练好的机器学习模型对处理完毕的 MFCC 特征进行拟合，计算出不同关键词的可能概率。之后根据概率判断出这时的关键词。

传统机器学习模型在进行识别时，会根据输入特征的不同产生出若干个概率，分别对应不同的输出标签。之后模型从中挑选出最大的一个，作为机器学习模型的实际输出。这种方法在大多数情况下被证明是一个优秀的方法，不过依旧存在着一个问题。在传统机器学习模型进行识别时，通常认为连续的前后两个输入之间是完全独立没有关系的信息，可以不考虑两者之间的联系进行独立的识别。不过对于本系统来说，连续输入的两个要进行识别的音频是有极其密切的关系的。两者之间有 96%的部分都是完全相同的，这也就要求了这两个输入产生的输出之间应该也存在着一定的联系，产生一个输出时应该考虑到相邻语音分片产生的输出的影响。但是传统的机器学习模型不是这么工作的。对于连续两个输入来说，其输出结果互相独立，没有影响。这样的计算会丧失掉一部分进行语音分片工作的意义。

如果想在进行输出时考虑到之前的识别信息，进行传统上输出识别标签值是不可行的。因为标签值作为离散信息，在识别过程中不包含任何有用的信息，难以利用。考虑到这一点，本系统使用的机器学习模型采用概率值作为识别结果输出而不是标签值作为识别结果输出。关键词识别部分对输入的语音信号进行识别

后，会产生识别为每个关键词的概率。将若干个连续分片中的同一个关键词识别概率结果进行相加，用这个和值来进行关键词判决，之后产生识别的标签值。这样可以保证关键词输出的平稳性，提高识别效果。考虑到对实时语音信号进行分片时步长很小，可以认为若干个连续的语音分片的识别结果应该是稳定的。如果此时处于一个关键词语音的范围内，关键词识别部分输出的概率值也应该是平稳的，在若干个连续分片内变化不大。考虑到这一点，系统对在这里人为设定一个用来衡量概率的阈值。如果和值超过设定的阈值，那么系统认为识别到了关键词，将这个和值对应的关键词输出。

同时之所以采用计算出关键词概率并且再进行一次判断来进行关键词识别而不是直接计算出关键词，还有一个原因。因为在实时关键词识别中绝大部分的语音信号都不包含我们需要识别的关键词信息。所以使用关键词的识别概率判断时，那些非关键词信息就会因为任何一个关键词的概率和达不到阈值门限而无法输出特定的关键词被认为是无关信息。此外，采用关键词的识别概率相加的方式可以避免在某一个语音分片中语音信号突然波动带来的影响，使输出更加平稳，提升识别的精确率。

这时存在一个灯光设备的占用问题。灯光设备在进行灯光动作的播放时会占用一定的时间。有时候输入的关键词过密，同时灯光动画时间比较长的话，会导致在前一个动画没有播放完成时，下一个动画因为灯光设备被占用而导致无法播放，使系统出现问题。为了应对这种情况，本模块增加了一个灯光设备占用记录部分。在关键词识别开始时，将占用标记置为 0，意味着这时灯光设备空闲。之后在每次输入一个语音分片后，都读取当前时间并进行占用标记判断，如果此时标记为 0，则可以进行关键词识别；如果此时占用标记为 1 且当前时间距离上次标记置为 1 的时间之差超过了规定阈值，则认为上一个灯光动画已经播放完成，可以进行关键词识别，并将标记置为 0；如果此时占用标记为 1 且当前时间距离上次标记置为 1 的时间之差没有超过规定阈值，则认为上一个灯光动画还没有播放完成，不可以进行关键词识别，直接将这个语音分片删除。在识别出关键词后，将占用标记置为 1，意味着这时灯光设备被占用，并记录下此时的时间，作为灯光动画开始时间。增加的灯光设备占用记录部分，可以有效避免因为灯光设备的占用导致系统堵塞的风险，减少了资源的开销。这个灯光占用时间参考过实际使用的灯光动作之后 设置为 4s。

### 3.5.3 指令转化

经过处理的实时语音信号输入到机器学习模型后，模型会对每个关键词产生

出识别概率，并根据概率识别出关键词，并生成对应的指令信号。本部分采用 Python 语言程序对该模块进行编程，其中算法 2 与算法 3 需要同时运行，伪代码如表 3-5 和表 3-6 所示：

表 3-5 实时语音信号采集算法

Table 3-5 Real-time voice signal acquisition algorithm

| 算法 2. 实时语音信号采集 |
| --- |
| 输入：实时语音信号 |
| 输出：音频文件 |
| 1. 将采集到的实时语音信号转化为语音流 |
| 2. chunk←320 |
| 3. for i from 0 to 10000 |
| 4. $data_i$ ← chunk 长度的语音流 |
| 5. frame[i] ← $data_i$ |
| 6. $file[]_i$ ← frame[1], frame[2],⋯,frame[100] |
| 7. frame[] ← frame[2:] |
| 8. end for |

表 3-6 关键词识别算法

Table 3-6 Keyword spotting algorithm

| 算法 3. 关键词识别 |
| --- |
| 输入：音频文件 |
| 输出：灯光指令 |
| 1. for i from 0 to 10000 |
| 2. now_time← 当前时间 |
| 3. if now_time - occupy_time<3 then |
| 4. occupy ← 1 |
| 5. occupy ← 0 |
| 6. end if |
| 7. if occupy=1 then |
| 8. 删除$file[]_i$ |
| 9. 从$file[]_i$中提取语音信号$sig_i$ |
| 10. 将$sig_i$中小于阈值的元素置零 |
| 11. 在$sig_i$列表右侧增加 0 元素，补充到规定长度 |

表 3-6 关键词识别算法（续）

Table 3-6 Keyword spotting algorithm(extend)

12. 对$sig_i$进行预处理并提取$mfcc_i$
13. 使用机器学习模型对$mfcc_i$进行预测
14. $ans[]_i$ ← 识别概率
15. $a[]_i$ ← 连续 5 个$ans[]_i$对应项之和
16. if max$a[]_i$>阈值 then
17. order ← 对应的关键词
18. occupy ← 1
19. occupy_time ← 当前时间
20. 将 order 传递给灯光设备
21. 删除$file[]_i$
22. end if
23. end if
24. end for

## 3.6 灯光设备控制

该模块用于控制灯光设备，包括以下三步：灯光动作的设计，串口通信，灯光动作的执行。系统将预先设计好的灯光动作保存在灯光动作库中。当 DMX512 设备通过串口通信接收到实时语音识别模块发送到的指令信号后，读取灯光动作库，根据指令信号检索出对应的灯光动作，生成灯光设备的通道控制信号，并通过串口通信将其发送给灯光设备。灯光设备接收到灯光设备控制信号后做出对应的灯光动作。灯光设备控制模块流程图如图 3-9 所示：

### 3.6.1 灯光动作的设计

如今大量的灯光设备采用广泛应用的 LED（发光二极管）作为灯具。本系统使用的灯光设备包含 7 颗 10W 红绿蓝黄四色合一大功率 LED，可以完成指定图案与灯光动作的设计。

本系统中使用的灯光设备由 DMX512 通道控制，在不同的通道填写不同的值可以改变灯光的参数。本模块主要使用的通道可以控制灯光图案在 X 轴、Y 轴进行移动，改变灯光图案的形状、大小、颜色，使灯光图案进行平移。在这里在 Visual Studio 2019 平台上使用 C++语言对其进行编程实现。实现方法如下：首先

读取灯光设备的接口并检测其连接状态，如果连接正常的话，将编写好的若干组数组作为 DMX512 通道参数输入。灯光设备接收到这若干组通道参数，并将这些通道参数置入对应的灯光通道中，这样就可以实现灯光控制。

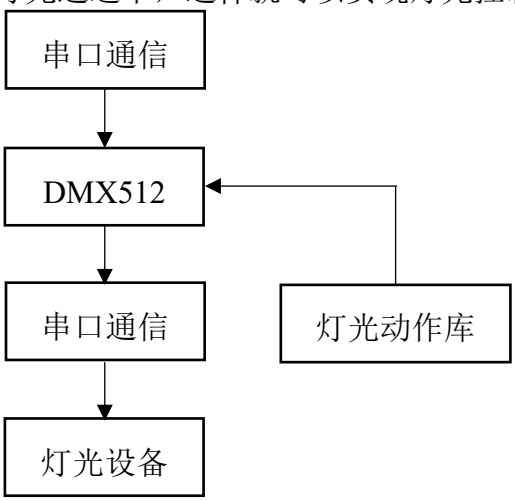


图 3-9 灯光设备控制模块流程图

Figure 3-9 Flow chart of lighting equipment control module

但是对于本灯光设备，同一时间只能输出一个图案，如果想实现更加复杂的情形的话，需要同时显示出多个图案。由于人眼有视觉暂留，在人眼所看到的图案消失之后，人眼依然可以在大脑里保留其印象长达 0.1-0.4 秒。电影在播放时，每秒钟放送 24 张画面，就可以显示出连续的影像，用的就是这个原理。这里使用循环语句，交替切换两个不同的图案，每次图案只停留 0.05 秒，在人眼看起来就像是同时演示了两种不同的图案，分别做不同的动作，就可以实现两个图案同时播放的效果。

在之前的模块一共设计了四个关键词，分别是：“红方朝”、“红方向”、“双方在”以及“蓝方攻”，分别意味着“红方朝一个地方移动”、“红方向安全的地方转移”、“双方在初始的位置驻扎”以及“蓝方攻击红方”。每个关键词对应一个独特的灯光动作，如表 3-7 所示。

表 3-7 关键词与灯光动作对应

Table 3-7 Keywords corresponding to light actions

| 关键词 | 灯光动作 |
|---|---|
| 红方朝 | 红色圆圈斜向移动 |
| 红方向 | 红色圆圈水平地移动，蓝色圆圈在向红色圆圈初始位置移动 |
| 双方在 | 同时显示红色圆圈和蓝色圆圈，两者在初始位置闪烁 |
| 蓝方攻 | 蓝色圆圈变为蓝色箭头，向红色圆圈移动 |

使用 C++语言程序对这些灯光动作进行编程，伪代码如表 3-8 所示：

表 3-8 灯光动作设计

Table 3-8 Lighting action design

| 算法 3. 灯光动作的设计 |
| --- |
| 输出：灯光动作可执行文件 |
| 1. 读取串口编号 |
| 2. 打开灯光设备 |
| 3. for t from 0 to 100 |
| 4. $b_1^t, b_2^t, \dots, b_n^t$为灯光通道参数 |
| 5. $buf_t[] \leftarrow b_1^t, b_2^t, \dots, b_n^t$ |
| 6. 将$buf_t[]$导入灯光设备中 |
| 7. 休眠，等待灯光动作完成 |
| 8. end for |
| 9. 关闭 |

### 3.6.2 串口通信

串口通信模块用于在计算机主机，DMX512 设备，灯光设备之间通信。计算机主机将指令信号通过串口通信传递给 DMX512 设备，DMX512 设备接收到这个指令信号，并对其进行频率调制与信号增强，之后将其转化为通道信息并通过串口通信传递给灯光设备。

串口通信模块工作时首先需要读取配置文件中预先写好的 DMX512 设备对应的端口号，然后根据端口号通过串口通信将打开 DMX512 设备并读取它的工作状态，并且设置好读取模式、接收待机数据的时间以及缓存空间。DMX512 设备保持接收状态，一旦接收到计算机主机发送过来的指令信号，便将其存入缓存空间，并对应转化为将其转化为 DMX512 数据包。数据包内包含有用来控制灯光设备的各个通道信息。灯光设备接收到这个数据包之后读取里面包含的通道信息，并调整对应的通道参数，做出规定的灯光动作。

### 3.6.3 灯光动作的执行

对于编译好的 C++语言程序，系统将其保存到本地的灯光数据中，记录下灯

光动作与关键词的对应关系。当灯光设备控制模块接收到灯光指令时，按照对应关系启动该 C++语言程序，就可以执行对应的灯光动作。灯光效果如图 3-10 所示：

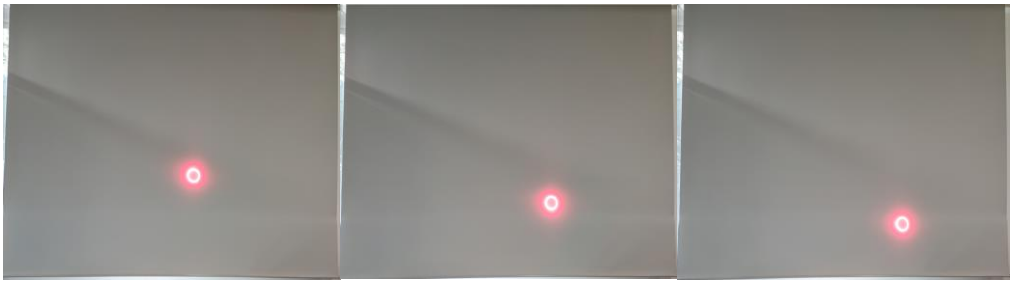

图 3-10 灯光效果图

Figure 3-10 Photo of lighting effect

这个效果图代表的灯光动作关键词为“红方朝”，转化为的灯光动作就是一个红色圆圈从左上角慢慢移动到右下角。

## 3.7 编程实现与工作量

针对本文提出的实时关键词识别的智能实体沙盘灯光控制系统，本人通过编程来实现其各个模块的功能。编程使用了 Python 语言与 C++语言。Python 语言包含丰富的第三方库资源，可以很好地完成自然语音处理以及机器学习模型构建的工作。C++语言相较于 Python 语言更接近底层，可以更好地对硬件设备进行操控。在程序编写过程中，本人完成的代码量为 400 行左右。下面对每个模块实现的方法进行叙述。

第一部分为音频数据获取模块，本人使用百度语音合成平台对关键词数据进行获取。百度语音合成平台提供 API 接口，可以供研发者调用。本人通过输入关键词文本来获取可以播放的音频文件，将这些音频文件作为语音数据集保存在本地。调用平台共获得了 2400 条含有关键词的语音文件与 600 条非关键词的噪声文件。

第二部分为机器学习模型训练模块。Python 语言为语音特征提取与机器学习模型构建提供了开源库供用户使用。本人对输入的音频文件进行了预处理与 MFCC 特征提取等操作，并比较了不同机器模型对关键词识别的不同评价指标，选择了合适的机器学习模型并对其超参数进行了调整。

第三部分为实时关键词识别模块。这部分除了包含机器学习模型训练模块中所涉及的工作之外，还包含了本人编写的实时语音分片部分，灯光设备占用检测部分，概率判断输出部分，指令转化部分以及时间对齐部分。可以对实时语音进

行处理。

第四部分为灯光设备控制模块。这部分采用 C++语言编写，编写了灯光设备的控制代码，并为每个关键词配上了对应的动作，配置了使用串口通信与灯光设备之间通信的方法。

## 3.8 本章小结

本章叙述了本文提出的实时关键词识别的智能实体沙盘灯光控制系统的整体设计思路与实验流程，介绍了音频数据获取模块，机器学习模型训练模块、实时关键词识别模块以及灯光设备控制模块四个模块的原理、作用、流程以及实现方法。介绍了编程实现的方法以及对工作量进行的总结。

# 4 系统评估

本章对上文所提出的实时关键词识别的智能实体沙盘灯光控制系统进行了试验评估。首先对评价方法进行了介绍，交代了系统工作的环境，介绍了评价的指标并且描述了评价流程。之后对实验结果进行了分析，评估了不同机器学习模型在本系统中的表现。

## 4.1 评价方法

### 4.1.1 软硬件环境

本文在同一台电脑上进行了机器学习模型训练与灯光控制工作。软硬件环境如下：

处理器：Intel(R)Core(TM)i7-8550U CPU@1.80GHZ 2.00GHZ

内存(RAM)：16.0GB

系统类型： 64 位 windows 操作系统

软件环境： visual studio 2019 ， Jupyter notebook

网络环境： 校园网

DMX512 设备：FQSD512-PR

灯光设备：7 颗 10W 红绿蓝黄四色合一大功率 LED

硬件设备图片如图 4-1，图 4-2 所示：


图 4-1 DMX512 设备

Figure 4-1 DMX512 device

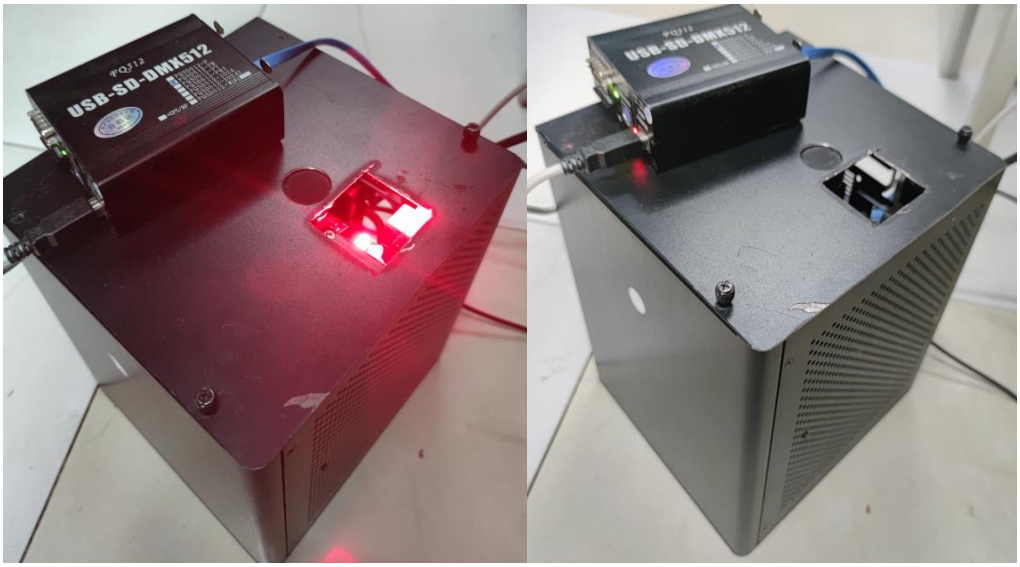


图 4-2 灯光设备

Figure 4-2 Lighting equipment

### 4.1.2 评价标准

在二元分类中可以把样本分为正样本与负样本，各代表一种样本。识别结果划分为正类(Positive)与负类(Negative)。正类指被识别为正的样本，负类指没有被识别为正的样本。可以把识别对错划分为真(True)与假(False)。真指识别正确，假指识别错误。对这四种情况两两组合可以产生四种实际情况，分别叫做：TP,FP,TN,FN。TP 表示正样本被识别为正类，识别正确。FP 表示负样本被识别为正类，识别错误。TN 表示负样本被识别为负类，识别正确。FN 表示正样本被识别为负类，识别错误。这四种实际情况将用在以下评价指标的计算中。这四种实际情况可以被画为 2*2 的混淆矩阵，如表 4-1 所示：

表 4-1 混淆矩阵

Table 4-1 Confusion matrix

| | | 真实值 | |
|---|---|---|---|
| | | P | N |
| 预测值 | P’ | TP（真阳性） | FP（假阳性） |
| | N’ | FN（假阴性） | TN（真阴性） |

对于机器学习系统的评价指标可以很容易想到使用准确率(Accuracy)来衡量，其定义如下：

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4-1}$$

可以看到，准确率表示的是所有预测正确的样本数量与全部样本数量的比值。准确率虽然可以判断系统的识别效果，不过在样本数偏差较大的情况下并不能很好地对系统的结果进行衡量。比如全部样本有 50 个，其中有 45 个正样本，5 个负样本，样本严重失衡。那么模型如果把所有样本都预测为正样本，就可以获得 90%的准确率，这样显然是不能体现到预测的作用。

为了解决准确率无法有效地衡量预测结果的问题，研究人员引入了精确率(Precision)和召回率(Recall)来对机器学习模型的识别效果进行衡量，其定义如下：

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4-2}$$

$$Recall = \frac{TP}{TP+FN} \tag{4-3}$$

精确率表示的是正样本被识别为正类的数量与所有被识别为正类的样本的比值。召回率表示的是正样本被识别为正类的数量与所有正样本的比值。精确率和召回率从不同的角度反映了识别的准确性，不过精确率和召回率在实际评估使用中很难同向变化，召回率的增加一般会导致精确率的下降。如果要综合考虑这两个指标，可以使用综合评价指标$F_1$，其定义如下：

$$F_1 = \frac{2*\text{Precision}*Recall}{\text{Precision}+Recall} \tag{4-4}$$

$F_1$为精确率与召回率的调和平均数，可以有效地描述机器学习模型的识别效果。当$F_1$较高时，说明系统有效性比较高。

对于本系统来说，由于有四个关键词，并不能算是一个传统的二分类问题。不过在对每一个关键词进行评价时，都可以把该关键词当作正样本，把被识别为该关键词的情况当作正例，被识别为其他关键词的情况或者没有被识别出来的情况当作负例。可以对每种关键词计算其的精确率以及召回率，之后对其计算平均值就可以用来估计该系统的性能。准确率由系统整体中所有预测正确的样本数量与全部样本数量的比值计算的到。$F_1$值则使用系统的平均精确率以及平均召回率计算得到。

对于实时关键词识别，还有一个重要指标，就是识别速度。实时关键词识别要求在讲话人说话的同时就进行关键词识别，要求有较强的实时性，一旦识别速度无法满足要求，还会出现较大的延迟，更有甚者会导致输出的错位。因此，识别速度过慢会影响识别效果以及观看体验。同时，识别速度只要满足一定的门限，可以正好处理输入的语音就可以，识别速度过快也会导致对资源的浪费与识别准确度。考虑到实时关键词识别与对测试集进行测试时实际系统环境是不同的，第三章中对不同模型的识别速度评估不能完全描述实际情况，这里还需要对实际使

用中的识别速度进行评估。

### 4.1.3 评价步骤

首先设计若干段包含关键词的文本，使用百度语音合成平台将这些文本转化为音频文件。由另外一个设备进行播放，这样就能模拟讲解员发出的实时语音信号，也能防止计算机主机因为自我屏蔽不能正确记录下本机发出的声音。这时，计算机主机同时启动实时语音识别模块以及灯光设备控制模块，通过话筒接收到播放的实时语音流，同时对实时语音流进行处理，交给机器学习模型进行识别，如果识别到关键词，就向灯光设备模块发送一个控制信号，使灯光设备做出指定的动作。

需要记录下来用来评价的参数有：原始音频中每个关键词的数目与出现的时间，识别结果中不同时间所识别出来的关键词内容。对于每一种关键词，如果在其出现的时间范围内被正确地识别出来，同时灯光设备做出了指定的动作，就视作识别正确。如果在其出现的时间范围内被识别为其他关键词，就视作识别错误。

同时为了对实时性进行评价，需要记录下系统在识别一段实时语音信号所需花费的时间，用来与这段实时语音信号的总时间作对比。此外，为了对延迟进行评价，还需要记录下从识别到关键词到灯光设备做出对应的灯光动作之间的时间。这里读取这两个模块的系统时间，计算上述两个时间点之间的系统时间之差，就可以计算出该部分的系统延迟。

## 4.2 实验结果分析

本系统在音频数据获取模块中使用了百度语音合成平台合成了 40 段长度各为 100 秒的音频文件，每个音频文件包含 4 种关键词，每种 2 个，总计 320 个关键词。使用另外一个设备进行播放，计算机主机使用话筒接受这个外部语音信号，用来模拟系统识别讲解员发出的实时信号的场景。之所以不使用计算机主机进行播放的原因是因为计算机主机在录音时会屏蔽本机发出的声音，防止本机噪声干扰，而本机发出的实时语音也同时被抑制，无法被话筒录制。下面使用了上文中表现较好的决策树模型与逻辑回归模型进行测试，统计其的识别结果，并计算出其的评价指标并且评价其识别效果。

当在实时关键词识别模块中使用决策树模型作为机器学习模型时，各项指标如表 4-2 所示：

表 4-2 决策树模型的评价指标

Table 4-2 Evaluation indicators of decision tree model

| | 准确率 | 精确率 | 召回率 | F1 |
|---|---|---|---|---|
| 红方朝 | -- | 0.688 | 0.825 | -- |
| 红方向 | -- | 0.750 | 0.750 | -- |
| 蓝方攻 | -- | 0.614 | 0.875 | -- |
| 双方在 | -- | 1.000 | 0.375 | -- |
| 平均值 | 0.710 | 0.763 | 0.706 | 0.733 |

当在实时关键词识别模块中使用逻辑回归模型作为机器学习模型时，各项指标如表 4-3 所示：

表 4-3 逻辑回归模型的评价指标

Table 4-3 Evaluation index of logistic regression model

| | 准确率 | 精确率 | 召回率 | F1 |
|---|---|---|---|---|
| 红方朝 | -- | 0.952 | 0.500 | -- |
| 红方向 | -- | 0.794 | 0.775 | -- |
| 蓝方攻 | -- | 0.658 | 0.650 | -- |
| 双方在 | -- | 0.529 | 0.800 | -- |
| 平均值 | 0.681 | 0.733 | 0.681 | 0.706 |

以上两张表格中左列为不同的关键词，表中数值代表对应关键词的评价指标的数值。“--”表示这个数值没有意义，因为准确率与$F_1$值都是对总体进行评价，不能对单个关键词进行评价。平均值一行中精确率与召回率是由四个关键词的对应指标做算术平均计算得到的，而准确率是由整体系统中的数据计算出来的，$F_1$值则是由平均精确率与平均召回率计算的到。

从以上两个模型的实验结果与评价指标可以看到，不同机器学习模型在实时语音识别模块中的识别的效果是不同的。决策树模型表现略优于逻辑回归模型。不过决策树模型容易没有正确识别出“双方在”而且容易把其他关键词错误地识别为“蓝方攻”。逻辑回归模型则容易没有正确识别出“红方朝”而且容易把其他关键词错误地识别为“蓝方攻”或“双方在”。这是由于训练集中的训练样本

无法满足机器学习模型要求的均匀分布导致的。这两个模型的准确率的平均值虽然在 0.7 左右，不过考虑到在实际情况中，讲解员发现在说出关键词后灯光设备没有做出规定的动作，这时还可以再重复一次关键词来进行补救。在这种情况下，在两次识别中至少有一次能识别对关键词的概率达到了 0.9，基本可以满足实际需求。

考虑到讲解员在讲解时语速会有变化，为了模拟这个情况，输入语音的语速变化时，两个模型的识别准确率变化如图所示：

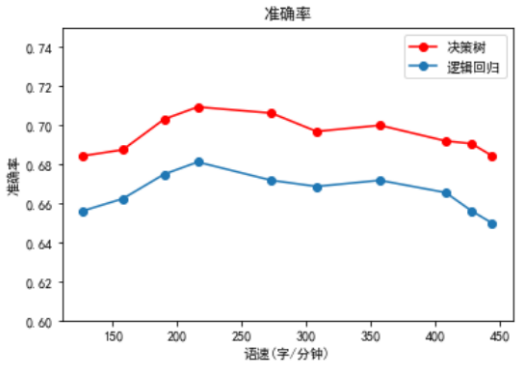


图 4-3 模型的准确率随语速变化关系图

Figure 4-3 The relationship between the accuracy of the model and the rate of speech

可以看到，两个模型在语速为 210 字/分钟的地方准确率最高，也是上文中测试集的语速。可以看到，模型的识别准确率在语速中等时略高，语速慢和语速快时略低，总体变化不大。这是由于在训练时使用的训练集中语速被设置为了中等，可以更好地模拟实际情况下人讲话的速度。

对于实时关键词识别系统比较关心的识别时间问题，本系统做得比较好。对于每个长度为 100 秒的实时语音，在播放的同时模块就开始对它进行识别，平均实时识别时间在 102 秒左右，存在着 2 秒的总延迟时间。音频文件时间长度的变化对这个总延迟时间的影响较小，两个模型的总延迟时间变化如下图所示：

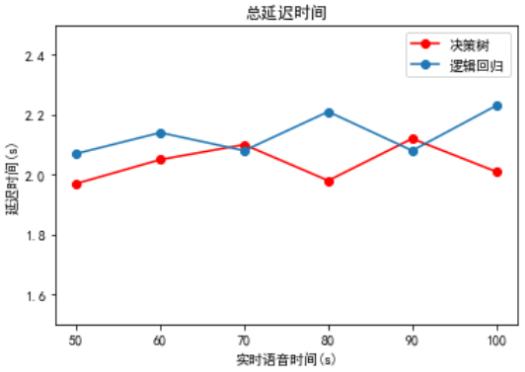


图 4-4 总延迟时间与总时间变化关系图

Figure 4-4 Relationship between total delay time and total time

可以看到，随着实时语音信号长度的变化，总延迟时间基本不变，为系统本身花费的时间。考虑到包含启动延迟等因素的干扰，可以说本系统能够在规定时间内完成对实时关键词的识别，具有良好的实时性。

在实时关键词识别模块识别到关键词之后，向灯光设备控制模块发送指令信号，调动灯光设备做出指定的动作。在灯光设备控制模块中，从将关键词识别出来之后到灯光设备响应做出灯光动作之间也必然存在一定的延迟，这个延迟大小也会一定程度上影响用户的观看体验。这部分的平均延迟为 2 秒，考虑到灯光动画并不是只能在关键词处这一个时间点播放，而是可在以关键词处周围数秒的一个时间段内播放，都可以表现出正确的效果，这个延迟并不会对系统的演示效果造成负面效果，可以满足系统要求。

## 4.3 本章小结

本章对第三章提出的实时关键词识别模块与灯光设备控住模块进行了测试与评估。其中对决策树模型与逻辑回归模型进行了实际效果的比较，对系统的实时性进行了评估。整体评估结果表明，使用本地的实时关键词识别系统来控制智能实体沙盘的灯光系统是可行的。同时采用决策树模型时系统的表现比采用逻辑回归模型时要更优。

# 5 结论

## 5.1 工作总结

智能实体沙盘在许多专业领域中有着重要的作用，对这些专业领域的从业者来说，总是希望智能实体沙盘的功能越多越好。因此在智能实体沙盘的发展过程中，设计者总是希望于给智能实体沙盘添加更多的软硬件设备，从而实现更加丰富的功能。

现如今的智能实体沙盘，在不依赖于外部网络的关键词识别功能上还有发展的空间。本文对智能实体沙盘增加了基于本地机器学习模型的关键词识别模块，可以实现基于关键词识别的灯光控制。本系统不依赖于外部网络，可以在本地做到关键词识别并控制灯光设备，使其实时做出指定的动作行为。其中关键词识别部分使用机器学习模型。因为机器学习模型消耗资源少，节约成本，并且识别速度比较快，比较符合本系统中的场景。

本系统的主要流程如下：在准备阶段，建立一个关键词库，设计每个关键词对应的灯光动作并将其保存在本地的灯光动作库中，通过网络上的语音合成平台获取关键词对应的音频文件，从中提取特征并训练机器学习模型。在识别阶段，系统对实时输入的语音信息进行分片与特征提取，在机器学习模型中进行识别，之后将识别结果转化为控制指令并发送给 DMX512 控制器。在灯光执行阶段，DMX512 控制器接收到这个控制指令，并控制灯光设备做出规定的灯光动作。

本文的独特之处有以下两点：

一是为智能实体沙盘系统增添了基于本地关键词识别模块的关键词识别部分，可以通过对讲解员发出的语音关键词在本地进行实时识别，从而控制灯光设备。目前在智能实体沙盘中应用的语音识别或者关键词识别模块都是基于网络语音识别平台的，本系统的出现改变了这一状态。

二是将机器学习模型应用在了本系统的关键词识别中。目前的关键词识别系统多采用神经网络模型，少部分采用支持向量机，资源开销大，无法应用在本系统这样算力无法支持神经网络开销的系统中。本系统采用决策树模型与逻辑回归模型作为机器学习模型对关键词进行识别，丰富了对关键词识别的途径。

## 5.2 未来工作展望

本文提出了一个实时关键词识别的智能实体沙盘灯光控制系统，能够不依赖

于外界网络，在本地通过识别讲解员发出的关键词来控制灯光设备。同时对系统中使用的不同机器学习模型进行了评估，挑选出了一个最适合本系统的机器学习模型。

本系统未来可以从以下几个地方进行改进：

(1)可以将几个关键词组合为一条高级指令。本系统目前只能将关键词与灯光动作一一对应，欠缺一定的灵活性。如果可以用若干个关键词对应一个灯光动作，将会进一步丰富系统的灵活性，提高用户的观看体验。不过更加丰富的灯光动作需要使用更多的灯光设备来实现。

(2)可以对灯光设备进行优化。目前的系统只使用了一个灯光设备，可以实现的灯光动作不够丰富。未来可以增加灯光设备，从而实现更加丰富的灯光动作，因而支持更多的关键词。不过如何协调多个灯光设备使其时钟对齐将会是一个新的问题。

(3)可以对智能实体沙盘的其他部分进行控制。智能实体沙盘包含多个模块，可以使用丰富的多媒体设备向观众进行展示。使用实时关键词识别方法不仅可以控制灯光设备，使用相同的方法可以对智能实体沙盘的其他部分进行控制，使系统更加灵活。

# 参考文献

[1] 韩晴.ERP 沙盘在企业培训中的应用研究[J].财讯,2016,(25):67-68.

[2] 戴辉. 浅谈沙盘技术在消防战术演练及实战中的应用[C].//2011 年度灭火与应急救援技术学术研讨会. 2011:392-394.

[3] 狄海廷,董希斌,肖生苓,等.森工立体程控沙盘虚拟实验室建设探讨[J].实验室研究与探索,2015,34(4):246-248,258.DOI:10.3969/j.issn.1006-7167.2015.04.064.

[4] 陆定中,李立明.轨道交通沙盘模型的操作显示界面设计研究[J].现代职业教育,2020,(45):166-167.

[5] 毕力格,达布希拉图,苏立娟,等.基于三维 GIS 的内蒙古人工影响天气电子沙盘系统设计及应用[J].气象科技,2018,46(1):207-213. DOI:10.19517/j.1671-6345.20170150.

[6] 黄剑飞.三维电子沙盘系统的研究与实现[D].湖南:湖南大学,2013. DOI:10.7666/d.Y2523310.

[7] 王峰,薛智勇,尚亮,等.多媒体互动沙盘模型电路控制系统研制[J].测绘技术装备,2013,(1):62-64.DOI:10.3969/j.issn.1674-4950.2013.01.021.

[8] 杜一川.小区域智能交通模拟系统演示平台[D].浙江:浙江大学,2006.

[9] 闫保中,杨欣颖. 基于 VC++ 的地铁智能实体沙盘控制系统设计[J]. 应用科技,2009,36(10):34-37,43. DOI:10.3969/j.issn.1009-671X.2009.10.010.

[10] 马蓉,赵九思. 基于 RFID 技术的智能停车系统模拟沙盘研究与设计[J]. 信息通信,2014(10):66-66. DOI:10.3969/j.issn.1673-1131.2014.10.038.

[11] 刘浩,李思其. 智能交通沙盘中的关键技术研究[J]. 电脑知识与技术,2018,14(34):184-185.

[12] 纪显俐,丁宇龙,杨双华.智能实体沙盘地理教学演示系统[J].系统仿真学报,2019,31(12):2816-2828.DOI:10.16182/j.issn1004731x.joss.19-FZ0407.

[13] 李莉,李韶军,智博远,等.快递运输智能监控沙盘在《物联网应用系统安装调试》课程教学改革探索与应用[J].电子世界,2020,(20):72-73.

[14] Q. Li, H. Liu and Z. Wu, “A Situation Visualization System Based on 2D&3D Electronic Sand Table, ”2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018, pp. 2057-2060, doi: 10.1109/IAEAC.2018.8577668.

[15] G. Huang and R. Guo, "Intelligent manor sand table demo system based on Internet of things and virtual reality technology," 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), 2017, pp. 571-575, doi: 10.1109/ITOEC.2017.8122361.

[16] P. Frantis, N. Kloudova and T. Klouda, "Technical issues of Virtual Sand Table system," 2017

International Conference on Military Technologies (ICMT), 2017, pp. 410-413, doi: 10.1109/MILTECHS.2017.7988794.

[17] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," in IEEE Signal Processing Letters, vol. 13, no. 1, pp. 52-55, Jan. 2006, doi: 10.1109/LSP.2005.860538.

[18] S.S.,S.and Volkman J． The relation of pitch to frequency． in American Journal of Psychology．1940．

[19] Schroeder,M ． ,Recognition of complex acoustic signals． Life Science Research Reports,1977．55:P．323—328.

[20] Ivica Rogina,Patrick Roessler．Automatic Speech Recognition．CMU&IRA.1998.

[21] Bridle, J.S., An Efficient Elastic-Template Method for Detecting Given Words in Running Speech, Brit. Acoust [J]. Soc. Meeting, ICASSP,1973:12-14.

[22] R. Christiansen and C. Rushforth, "Detecting and locating key words in continuous speech using linear predictive coding," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 5, pp. 361-367, October 1977, doi: 10.1109/TASSP.1977.1162983.

[23] C. Myers, L. Rabiner and A. Rosenberg, "An investigation of the use of dynamic time warping for word spotting and connected speech recognition," ICASSP '80. IEEE International Conference on Acoustics, Speech, and Signal Processing, Denver, CO, USA, 1980, pp. 173-177, doi: 10.1109/ICASSP.1980.1171067.

[24] A. Higgins and R. Wohlford, "Keyword recognition using template concatenation," ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, USA, 1985, pp. 1233-1236, doi: 10.1109/ICASSP.1985.1168253.

[25] J. G. Wilpon, C. H. Lee and L. R. Rabiner, "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech," International Conference on Acoustics, Speech, and Signal Processing,, Glasgow, UK, 1989, pp. 254-257 vol.1, doi: 10.1109/ICASSP.1989.266413.

[26] G. Hinton, L. Deng, D. Yu, and G. Dahl et al., “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Process.Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[27] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 4704-4708, doi: 10.1109/ICASSP.2015.7178863.

[28] A. L. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y.Ng, “Recurrent neural

networks for noise reduction in robust ASR," in Proc. Interspeech, 2012, pp. 22–25.

[29] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise," in Proc. ICASSP,2013, pp. 6822–6826.

[30] S. Kim, T. Hori and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 4835-4839, doi: 10.1109/ICASSP.2017.7953075.

[31] Alex Graves, Santiago Fernandez, Faustino Gomez, and J ´ urgen ¨ Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 369–376.

[32] Y. He, R. Prabhavalkar, K. Rao, W. Li, A. Bakhtin and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 2017, pp. 474-481, doi: 10.1109/ASRU.2017.8268974.

[33] 徐明星,郑方,吴文虎,方棣棠.连续关键词识别系统的拒识方法研究[J].清华大学学报(自然科学版),1998(S1):92-94.

[34] 袁长海,李星.基于关键词捕捉的中文语音网页浏览器[J].计算机工程与应用,2003(25):171-174+213.

[35] 米尔阿迪力江·麦麦提,吾守尔·斯拉木,努尔麦麦提·尤鲁瓦斯,热依曼·吐尔逊,艾尼宛尔·托乎提.基于智能手机的维吾尔语语音控制系统的开发[J].计算机应用与软件,2016,33(06):220-223+305.

[36] Vapnik, V.N. and Lerner, A.Y., 1963. Recognition of patterns with help of generalized portraits. Avtomat. i Telemekh, 24(6), pp.774-780.

[37] Y. Benayed, D. Fohr, J. P. Haton and G. Chollet, "Confidence measures for keyword spotting using support vector machines," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), Hong Kong, China, 2003, pp. I-I, doi: 10.1109/ICASSP.2003.1198849.

[38] Y. BenAyed, D. Fohr, J.P. Haton, and G. Chollet, "Keyword spotting using support vector machines," in 5th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 2002.

[39] Y. BenAyed, D. Fohr, J. P. Haton, and G. Chollet, "Recognition and rejection performance in word spotting systems using support vector machines," in 2nd WSEAS International

Conference on Signal, Speech and Image, Skiarhos Island, Greece, 2002.

[40] 陈太波,张翠芳.多特征和 SVM 改进的关键词识别系统[J].小型微型计算机系统,2019,40(11):2291-2296.

[41] 冯勇,周先军,盛秋林.一种基于 DMX 512 协议的照明控制系统[J].湖北工业大学学报,2018,33(2):59-61.

[42] 谭登峰,郑小玲,刘卫宏, 等.智能交互界面互联网在指挥调度平台中的应用[J].电视技术,2020,44(9):7-13. DOI:10.16280/j.videoe.2020.09.002.

[43] 刁学磊.智能实体沙盘人机交互系统研发[D].黑龙江:哈尔滨工业大学,2020.

[44] 郭永刚. 基于 STM32 的智能语音交互式沙盘控制系统设计与实现[D]. 甘肃:兰州大学,2017. DOI:10.7666/d.D01300002.

[45] 姚星辰,宋岩,安籽鹏. 三维电子沙盘系统的语音交互技术研究[J]. 数码设计（上）,2017,6(4):40-42. DOI:10.19551/j.cnki.issn1672-9129.2017.07.017.

[46] 邓金雪. 可编程 DMX 设备的控制设计和实现[D].北京:北京交通大学,2019.

# 作者简历及攻读硕士学位期间取得的研究成果

穆彦龙，男，1996 年 10 月生。2015 年 9 月至 2019 年 6 月就读于北京理工大学信息对抗专业，获得工学学士学位。2019 年 9 月至 2021 年 7 月就读于北京交通大学，取得工学硕士学位。

# 独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：  签字日期： 2021 年 5 月 31 日

# 学位论文数据集

表 1.1： 数据集页

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>智能实体沙盘，实时关键词识别，机器学习，灯光控制</td><td>公开</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">学位授予单位名称*</td><td>学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>工程</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名*</td><td colspan="2">并列题名</td><td>论文语种*</td></tr>
<tr><td colspan="2">实时关键词识别的智能实体沙盘灯光控制系统的研究</td><td colspan="2"></td><td>中文</td></tr>
<tr><td>作者姓名*</td><td colspan="2">穆彦龙</td><td>学号*</td><td>19125040</td></tr>
<tr><td colspan="2">培养单位名称*</td><td>培养单位代码*</td><td>培养单位地址</td><td>邮编</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>北京市海淀区西直门外上园村 3 号</td><td>100044</td></tr>
<tr><td colspan="2">学科专业*</td><td>研究方向*</td><td>学制*</td><td>学位授予年*</td></tr>
<tr><td colspan="2">电子与通信工程</td><td>信息系统</td><td>2</td><td>2021</td></tr>
<tr><td>论文提交日期*</td><td colspan="4">2021 年月日</td></tr>
<tr><td>导师姓名*</td><td colspan="2">李纯喜</td><td>职称*</td><td>副教授</td></tr>
<tr><td>评阅人</td><td colspan="2">答辩委员会主席*</td><td colspan="2">答辩委员会成员</td></tr>
<tr><td></td><td colspan="2">孙强</td><td colspan="2"></td></tr>
<tr><td colspan="5">电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ）<br>推荐格式：application/msword；application/pdf</td></tr>
<tr><td colspan="2">电子版论文出版（发布）者</td><td colspan="2">电子版论文出版（发布）地</td><td>权限声明</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td><td></td></tr>
<tr><td>论文总页数*</td><td colspan="4">56</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 21 项。</td></tr>
</table>