



浙江大學
ZHEJIANG UNIVERSITY

信号与系统课程设计报告

基于 mfcc 特征，vad 算法和机器学习的连续数字
语音识别系统

姓名 H-Hyunmin

学号

专业

院所

2024 年 6 月 28 日

目录

1	问题提出	3
2	MFCC 特征提取	3
3	语音活动检测 (VAD)	5
4	语音识别 (ASR)	7
5	连续数字语音识别系统	10
5.1	系统实现	10
5.2	代码仿真实验结果	10
6	总结	13

1 问题提出

在语音识别技术迅速发展的今天，语音识别系统已广泛应用于各个领域，包括语音助手、智能家居、自动客服等。尤其是连续数字串的语音识别在银行账户验证、电话号码输入和智能控制等方面具有重要应用价值。然而，如何有效地提取语音信号的特征并进行准确的识别仍然是一个具有挑战性的问题。

在信号实验课程中，我们已经实现了单个数字的语音识别。但是其中仍然存在识别不准确，泛化性不佳等问题。且如何实现单音频中连续数字的识别仍是一个问题。

在连续数字串的语音识别系统设计中，主要面临以下几个关键问题：

1. **语音特征提取**：语音信号具有非平稳性和复杂性，需要通过有效的特征提取方法将其转换为具有代表性的特征参数。梅尔频率倒谱系数（MFCC）是一种常用且有效的语音特征提取方法，能够较好地反映语音信号的频谱特征。

2. **语音活动检测（VAD）**：在连续语音信号中，需要准确地分离出语音段和非语音段。VAD 算法能够帮助系统有效地检测和分割语音信号，从而提高后续识别过程的效率和准确性。

3. **机器学习模型的选择与训练**：选择合适的机器学习算法，并对模型进行训练，以实现高效的语音识别。常用的模型包括隐马尔可夫模型（HMM）、深度神经网络（DNN）和卷积神经网络（CNN）等。这些模型需要大量的语音数据进行训练，以提高识别的准确率。

4. **连续数字串的识别**：在实际应用中，语音信号通常是连续的，包含多个数字串。如何在连续语音信号中准确地识别和分割每个数字，是实现高性能语音识别系统的一个难点。

基于以上问题，本报告提出了一种基于 MFCC 特征、VAD 算法和机器学习的连续数字串语音识别系统设计方案。该方案旨在通过有效的特征提取、准确的语音活动检测和高效的机器学习模型，实现对连续数字串语音信号的准确识别。本文将详细介绍系统的设计思路、关键技术和实现方法，并对实验结果进行分析和讨论。

2 MFCC 特征提取

梅尔频率倒谱系数（Mel Frequency Cepstral Coefficients，简称 MFCC）是语音信号处理中广泛应用的一种特征提取方法。它通过模拟人类听觉系统的感知方式，对语音信号进行分析和处理，从而获取更能反映语音本质特征的参数。以下简单介绍 MFCC 和实现 MFCC 特征提取过程。具体的代码实现可以参考仿照代码中的 mfcc 文件夹中的内容。

梅尔刻度 (Mel Scale)

人耳对频率的感知是 f 线性的，低频部分的分辨率较高，而高频部分的分辨率较低。梅尔刻度正是用来模拟这种非线性的感知方式。将实际频率转换为梅尔刻度的公式如下：

$$Mel = 2595 \times \log_{10}\left(1 + \frac{Frequency}{700}\right)$$

从梅尔刻度转换回实际频率的公式为：

$$Frequency = 700 \times (10^{\frac{Mel}{2595}} - 1)$$

梅尔刻度的引入，使得频率的变化更符合人耳的听觉特性，有效地模拟了听觉系统对声音的处理过程。

临界带 (Critical Band) 和 mel 滤波器

人耳对声音的感知不仅与频率有关，还与频率的分布密切相关。临界带将频率划分为若干个频带，每个频带对应一个滤波器。这些滤波器的集合被称为 Mel 滤波器组 (Mel filter bank)。研究表明，人耳对 200Hz 到 5kHz 范围内的语音信号最为敏感，这一频段对语音的清晰度有重要影响。

当两个不同响度的声音同时存在时，响度较大的声音会掩蔽响度较小的声音，这种现象被称为掩蔽效应。低频声音在内耳蜗基底膜上的行波传递距离大于高频声音，因此低频声音更容易掩蔽高频声音。基于这一特性，Mel 滤波器组设计为从低频到高频按临界带宽的大小由密到疏排列，对输入信号进行带通滤波。每个滤波器输出的信号能量作为基本特征，进一步处理后可作为语音识别的输入特征。

MFCC 的计算过程

MFCC 的计算过程包括以下几个步骤：

具体实现过程请查看仿真代码中的 `mfcc.ipynb`

1. **预加重 (Pre-emphasis)**：通过一个高通滤波器增强高频部分，补偿语音信号中高频成分的衰减。
2. **分帧 (Framing)**：将语音信号分成若干帧，每帧通常为 20-40 毫秒，并且相邻帧之间有一定的重叠。
3. **加窗 (Windowing)**：对每一帧信号乘以一个窗口函数（如汉明窗），以减少帧与帧之间的不连续性。
4. **傅里叶变换 (FFT)**：将每一帧信号从时域转换到频域，得到每帧信号的频谱。

5. **Mel 滤波器组 (Mel Filter Bank):** 将频谱通过 Mel 滤波器组, 得到每个滤波器的能量。
6. **取对数 (Logarithm):** 对滤波器输出的能量取对数, 模拟人耳对声音强度的感知。
7. **离散余弦变换 (DCT):** 对对数能量进行离散余弦变换, 得到 MFCC 系数。

3 语音活动检测 (VAD)

语音活动检测 (Voice Activity Detection, 简称 VAD) 是一种用于识别语音信号中的语音片段和非语音片段的技术。其主要目的是在连续的音频信号中区分出哪些部分是包含语音活动的, 从而过滤掉背景噪声和静音部分, 提高后续处理的效率和准确性。

VAD 算法广泛应用于各种语音处理系统中, 包括语音识别、语音增强、语音编码和语音合成等。通过准确地检测出语音活动段, 可以减少处理不必要的无用信息, 从而提高系统的性能和可靠性。

VAD 算法的实现

VAD 算法有多种实现方法, 每种方法都有其优缺点和适用场景。以下是几种常见的 VAD 算法实现方法:

1. **基于能量的 VAD:** 通过计算音频信号的短时能量来判断是否存在语音活动。此方法计算简单, 但在低信噪比环境下表现较差。
2. **基于谱熵的 VAD:** 通过计算音频信号每帧频谱的熵值来判断语音活动。该方法对包含大量噪声的音频信号有较好的鲁棒性, 但计算复杂度相对较高。
3. **基于自相关函数的 VAD:** 利用信号的自相关特性来区分语音和噪声。该方法能有效地区分语音信号和噪声信号, 但对非稳态噪声的处理效果有限。
4. **基于机器学习的 VAD:** 通过训练分类器 (如支持向量机、神经网络等) 来识别语音活动。该方法具有较高的准确性和鲁棒性, 但依赖于大量的训练数据和较高的计算资源。
5. **基于深度学习的 VAD:** 利用卷积神经网络 (CNN)、长短时记忆网络 (LSTM) 等深度学习模型, 捕捉更复杂的特征和语音活动模式。该方法在复杂噪声环境和多种语言的语音信号处理中表现出色, 但同样需要大量的训练数据和高性能计算资源。

基于 mfcc 特征提取和无监督机器学习的 VAD 算法

本设计使用基于 mfcc 特征提取和无监督机器学习（如 K-means 聚类）的 VAD 算法。其是一种结合了语音信号分析和机器学习技术的方法。该算法首先使用 MFCC 特征提取算法提取出语音信号的特征向量，然后使用无监督机器学习算法（如 K-means 聚类）对特征向量进行聚类分析，将包含语音活动的特征向量划分为一个聚类，将包含非语音活动的特征向量划分为另一个聚类。最后，根据聚类结果判断当前帧是否包含语音活动。

具体实现参考 vad 文件夹下的 py 程序

大致实现步骤如下：

1. 音频预处理：

- **加载音频文件：**读取音频文件并将其转换为适当的采样率。
- **噪声抑制：**对音频信号进行噪声抑制处理，以减少背景噪声的影响。

2. 特征提取：

- **短时傅里叶变换（STFT）：**将音频信号分割成短时帧，并对每帧进行傅里叶变换。
- **梅尔频率倒谱系数（MFCC）：**从每一帧的频谱中提取 MFCC 特征，这些特征模拟了人耳对声音的感知方式。

3. 聚类分析：

- **KMeans 聚类：**对提取的 MFCC 特征进行 KMeans 聚类，将特征分为语音和非语音两类。
- **高斯混合模型（GMM）聚类（可选）：**使用 GMM 对特征进行聚类分析，作为 KMeans 聚类的替代方法。

4. 语音段检测：

- **能量计算：**计算每个聚类类别的平均能量，用于区分语音段和非语音段。
- **标签分析：**通过分析聚类标签，识别出音频信号中的语音段。

5. 后处理：

- **合并相近的语音段：**将相隔较近的语音段合并，以避免将连续的语音段误认为多个独立的语音段。

6. 结果保存：

- **保存语音段信息：**将检测到的语音段的起始和结束时间索引保存到文件中，供后续分析和处理使用。

4 语音识别（ASR）

语音识别（Automatic Speech Recognition，简称 ASR）是指将人类的语音信号转换为相应的文本或指令的技术。随着计算机科学和人工智能的发展，语音识别技术在过去几十年里取得了显著进展，并在各个领域得到了广泛应用，如智能助手（如 Siri、Google Assistant）、语音翻译、语音控制的设备和应用（如智能家居、汽车语音控制系统）等。

语音识别的工作流程一般包括以下几个步骤：

1. **语音信号预处理**：对采集到的语音信号进行降噪、预加重、分帧、加窗等处理，以提高信号的质量和后续特征提取的效果。
2. **特征提取**：从预处理后的语音信号中提取出能够有效表征语音内容的特征参数，如梅尔频率倒谱系数（MFCC）、线性预测倒谱系数（LPCC）等。
3. **声学模型**：使用统计模型（如高斯混合模型，GMM）或深度学习模型（如卷积神经网络，CNN，或长短时记忆网络，LSTM）将提取的特征参数映射到音素或音素序列。
4. **语言模型**：根据语言的结构和统计规律，对识别出的音素序列进行解码，生成最有可能的词或句子。
5. **后处理**：对识别结果进行修正和优化，如纠错、标点符号添加等，以提高识别文本的准确性和可读性。

语音识别的实现

语音识别技术有多种实现方法，以下是几种常见的方法及其特点：

1. **基于模板匹配的方法**：这种方法将待识别的语音信号与预先存储的模板进行比较，以找到最匹配的模板。主要包括动态时间规整（Dynamic Time Warping，DTW）方法。此方法适用于小词汇量的应用，但对大词汇量和复杂语音信号的识别效果较差。
2. **基于隐马尔可夫模型（HMM）的方法**：隐马尔可夫模型是语音识别领域中最早广泛应用的统计模型。HMM 可以有效地处理语音信号的时序性和变异性，通过训练得到的 HMM 模型可以用于识别语音信号中的音素序列。该方法对中小型词汇量的语音识别有较好的效果。
3. **基于神经网络的方法**：随着深度学习的发展，基于神经网络的方法在语音识别中得到了广泛应用。主要包括卷积神经网络（CNN）、递归神经网络（RNN）、长短时记忆网络（LSTM）等。这些模型能够捕捉语音信号中的复杂特征和时序关系，大大提高了语音识别的准确率。特别是端到端（End-to-End）的深度学习模型，可以直接将语音信号映射到文本序列，简化了传统语音识别的复杂流程。

4. **基于 Transformer 的方法:** 近年来,Transformer 模型在自然语言处理领域取得了突破性的进展,也被引入到语音识别中。基于 Transformer 的模型,如 Transformers 和其变体(如 BERT、GPT-3),可以处理长距离的依赖关系和上下文信息,进一步提高了语音识别的效果。
5. **混合模型的方法:** 将不同模型的优势结合起来,如 HMM-DNN(深度神经网络与隐马尔可夫模型结合)方法,利用 HMM 处理时序信息,DNN 处理非线性特征。这种方法在大型语音识别系统中得到了成功应用。

基于卷积神经网络 (CNN) 的语音识别

我们的语音识别系统采用基于卷积神经网络 (CNN) 的语音识别系统的实现。该系统主要包括数据预处理、特征提取、模型构建、训练和推理等步骤。以下是对每个步骤的具体描述。具体的代码实现参考仿真代码。

数据集处理和特征提取

数据集来源于网络,将随机其分成训练集,验证集和测试集

音频文件通过 `scipy.io.wavfile` 库读取,并转换为适当的采样率。之后,使用 `python_speech_features` 库提取 MFCC (Mel 频率倒谱系数) 特征。这些特征能够有效地表征音频信号的频谱特性。

特征提取是将音频信号转换为模型能够处理的特征向量的过程。在本系统中,我们采用了 MFCC、以及其一阶和二阶差分特征。具体步骤如下:

1. 从音频信号中提取 MFCC 特征。
2. 计算 MFCC 特征的一阶差分和二阶差分。
3. 将 MFCC 特征及其差分特征拼接在一起,形成最终的特征矩阵。
4. 对特征矩阵进行截取或填充,使其形状一致。
5. 将特征矩阵转置并添加空维度,以符合 CNN 的输入要求。

模型构建

在模型构建部分,我们设计了一个卷积神经网络来处理音频特征。模型主要包括以下几层:

- **卷积层和批归一化层:** 网络包含多个卷积层,每层之后接一个批归一化层,以加速收敛和提高模型稳定性。

- **激活函数**：每个卷积层后应用 ReLU 激活函数，以引入非线性。
- **全连接层**：卷积层输出的特征经过全连接层进行分类。
- **Dropout 层**：在全连接层后添加 Dropout 层，以防止过拟合。

模型训练

模型训练过程使用 AdamW 优化器和交叉熵损失函数。为了提高模型性能和防止过拟合，训练过程中还采用了以下策略：

- **学习率调度**：使用 ReduceLROnPlateau 调度器，根据验证集精度自动调整学习率。
- **L2 正则化**：在损失函数中添加 L2 正则化项。
- **早期停止**：根据验证集精度保存最佳模型。

每个训练周期中，模型会在训练集上进行前向传播和反向传播，更新权重参数。之后，在验证集上评估模型性能，并根据验证结果调整学习率和保存最佳模型。

模型推理

模型推理包括在测试集上评估模型和对新音频进行识别。推理过程如下：

1. 加载预训练模型权重。
2. 将测试数据通过 DataLoader 加载到模型中进行推理。
3. 计算模型的预测精度，并输出预测结果。

推理过程中，模型会根据输入特征进行前向传播，输出预测的类别标签，并与真实标签进行比较，计算分类准确率。

模型效果

由于本设计仅为一个 demo，模型复杂度较低，数据集较小，因此模型效果可能无法达到实际应用的要求。但是，通过训练和测试，模型在测试集上的准确率基本达到了 90% 以上。

此外，模型的性能和泛化能力仍有不足，对于差异性较大的音频样本，模型可能无法准确识别，此外也可能存在一些过拟合的现象。因此，在实际应用中，需要进一步优化模型结构和参数，并使用更多的数据集进行训练和测试。

5 连续数字语音识别系统

5.1 系统实现

音频信号特征提取

我们对音频信号提取 mfcc 特征，作为 VAD 和机器学习的输入

语音活动检测（VAD）

语音活动检测（VAD）用于从音频信号中区分语音和非语音段，以得到数字语音段

卷积神经网络建模与训练

卷积神经网络用于对提取的 MFCC 特征进行分类，使用数据集进行训练

输入信号识别

系统对输入的音频信号进行下列处理

1. 使用 VAD 算法将输入音频分为多个语音段，每个段落对应一个数字语音
2. 对每段音频提取 mfcc 特征
3. 将每段音频的 mfcc 特征输入神经网络进行推理得到对应的结果

5.2 代码仿真实验结果

连续数字语音识别系统实现的代码位于项目 asr 文件夹下，具体实现参见源码。
下面简单展示系统实现效果：



图 1: 系统 GUI 界面

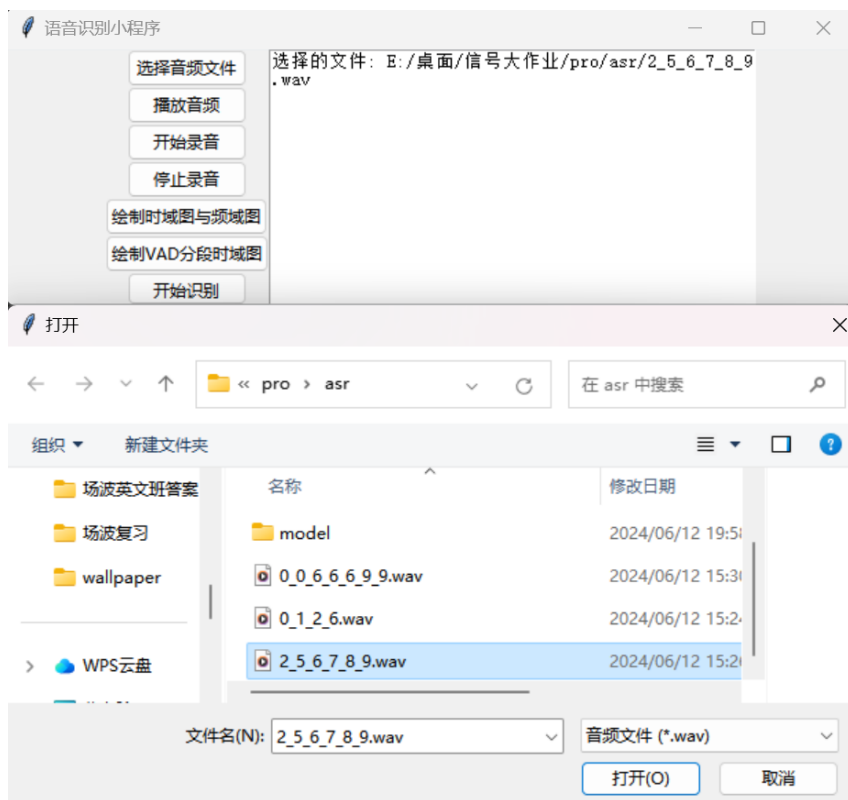


图 2: 选择音频

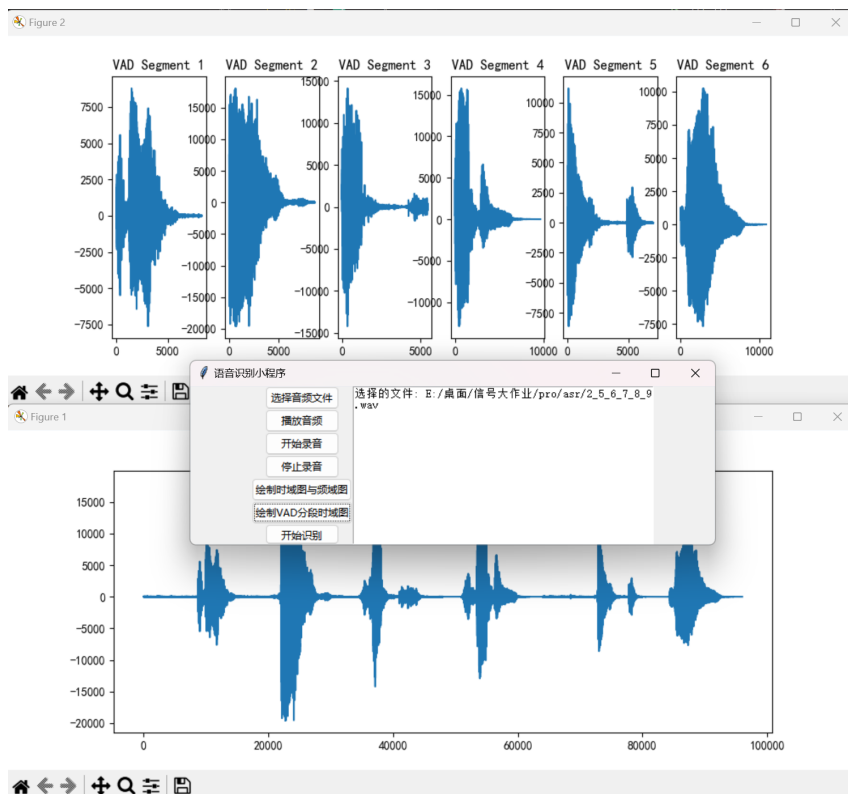


图 3: 绘制时域图和 VAD 识别结果

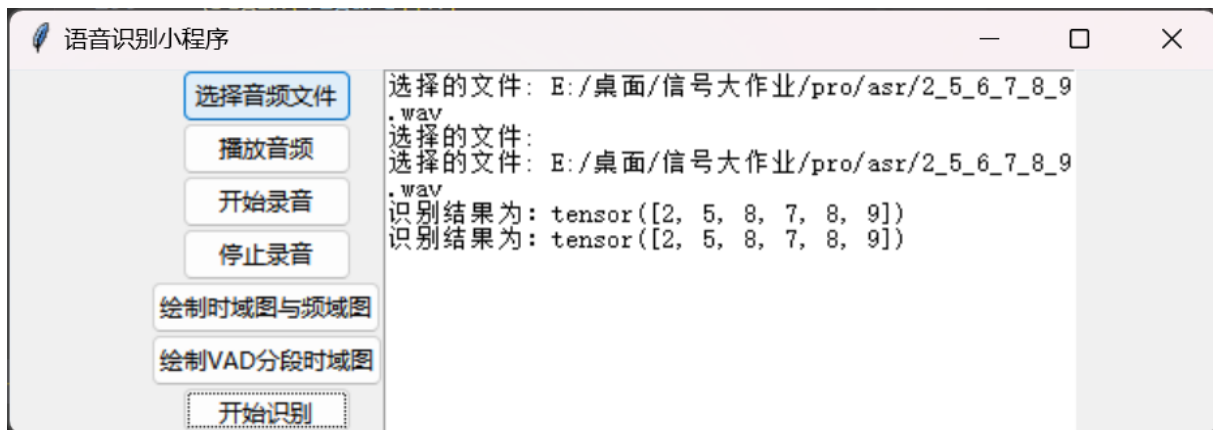


图 4: 数字音频识别结果

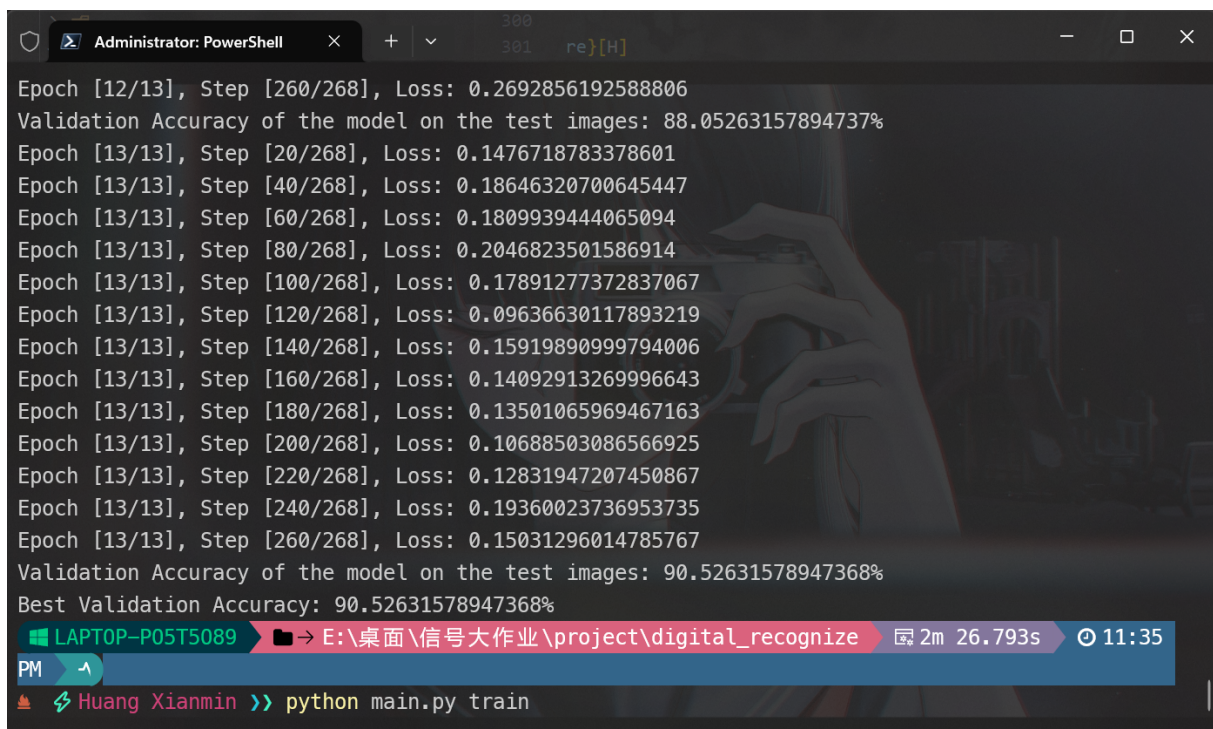
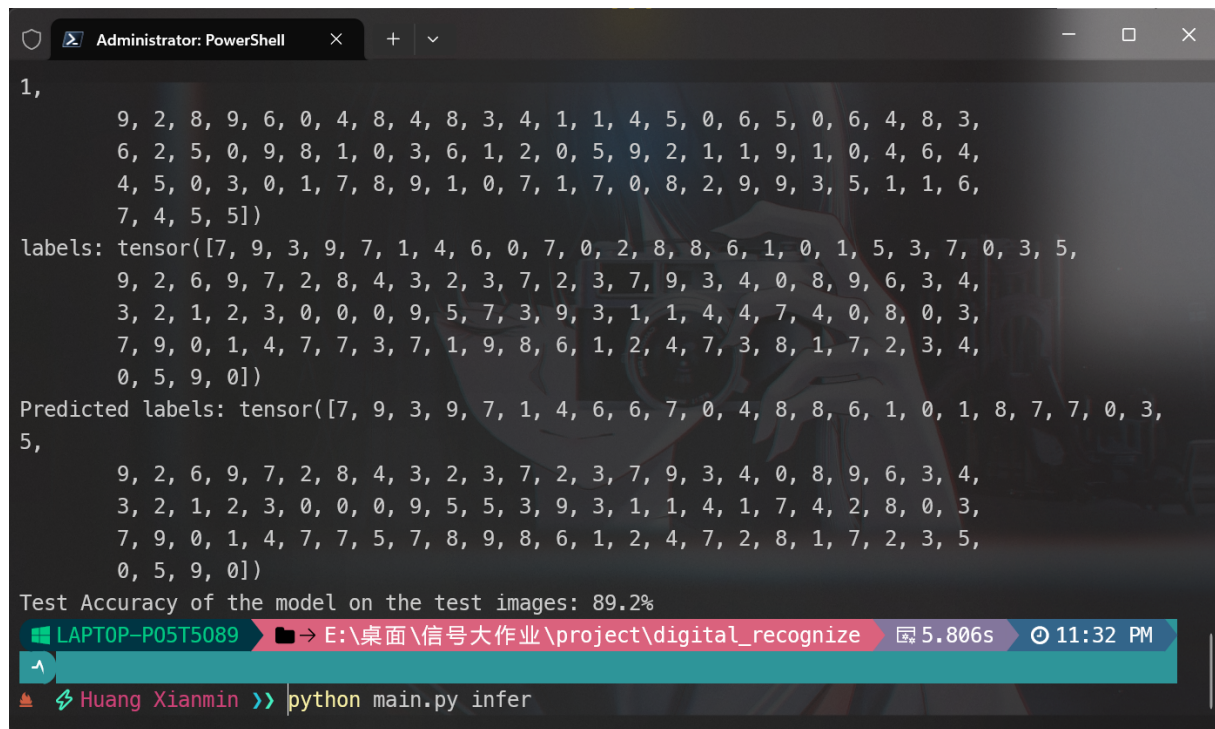


图 5: 模型训练



```
1,  
    9, 2, 8, 9, 6, 0, 4, 8, 4, 8, 3, 4, 1, 1, 4, 5, 0, 6, 5, 0, 6, 4, 8, 3,  
    6, 2, 5, 0, 9, 8, 1, 0, 3, 6, 1, 2, 0, 5, 9, 2, 1, 1, 9, 1, 0, 4, 6, 4,  
    4, 5, 0, 3, 0, 1, 7, 8, 9, 1, 0, 7, 1, 7, 0, 8, 2, 9, 9, 3, 5, 1, 1, 6,  
    7, 4, 5, 5])  
labels: tensor([7, 9, 3, 9, 7, 1, 4, 6, 0, 7, 0, 2, 8, 8, 6, 1, 0, 1, 5, 3, 7, 0, 3, 5,  
    9, 2, 6, 9, 7, 2, 8, 4, 3, 2, 3, 7, 2, 3, 7, 9, 3, 4, 0, 8, 9, 6, 3, 4,  
    3, 2, 1, 2, 3, 0, 0, 0, 9, 5, 7, 3, 9, 3, 1, 1, 4, 4, 7, 4, 0, 8, 0, 3,  
    7, 9, 0, 1, 4, 7, 7, 3, 7, 1, 9, 8, 6, 1, 2, 4, 7, 3, 8, 1, 7, 2, 3, 4,  
    0, 5, 9, 0])  
Predicted labels: tensor([7, 9, 3, 9, 7, 1, 4, 6, 6, 7, 0, 4, 8, 8, 6, 1, 0, 1, 8, 7, 7, 0, 3,  
5,  
    9, 2, 6, 9, 7, 2, 8, 4, 3, 2, 3, 7, 2, 3, 7, 9, 3, 4, 0, 8, 9, 6, 3, 4,  
    3, 2, 1, 2, 3, 0, 0, 0, 9, 5, 5, 3, 9, 3, 1, 1, 4, 1, 7, 4, 2, 8, 0, 3,  
    7, 9, 0, 1, 4, 7, 7, 5, 7, 8, 9, 8, 6, 1, 2, 4, 7, 2, 8, 1, 7, 2, 3, 5,  
    0, 5, 9, 0])  
Test Accuracy of the model on the test images: 89.2%  
LAPTOP-P05T5089 → E:\桌面\信号大作业\project\digital_recognize 5.806s 11:32 PM  
Huang Xianmin >> python main.py infer
```

图 6: 模型测试效果

6 总结

本报告提出并实现了一种基于 MFCC 特征、VAD 算法和卷积神经网络的连续数字串语音识别系统。该系统通过以下几个关键步骤实现了对连续数字串语音信号的有效识别:

1. **MFCC 特征提取:** 利用梅尔频率倒谱系数 (MFCC) 方法, 从语音信号中提取出具有代表性的特征参数, 使得语音信号的频谱特征得以充分体现。
2. **语音活动检测 (VAD):** 通过基于 MFCC 特征和无监督机器学习的 VAD 算法, 将连续的语音信号分割为语音段和非语音段, 从而提高了识别的准确性和效率。
3. **卷积神经网络建模与训练:** 设计并训练了一个卷积神经网络 (CNN), 用于对提取的 MFCC 特征进行分类, 从而实现了对单个数字的准确识别。
4. **连续数字串识别:** 结合 VAD 和 CNN, 对输入的连续数字串语音信号进行识别, 实现了对多个数字的连续识别。

通过上述方法, 本系统在连续数字串的语音识别任务中取得了较好的实验结果, 验证了设计方案的有效性。尽管当前系统的准确率和泛化能力在某些复杂情况下仍有待提

高，但本报告所提出的方法为进一步研究和改进语音识别技术提供了有价值的参考和基础。

未来的工作中，可以考虑以下几个方面的改进：

- **数据集扩展：**收集更多样化和大规模的语音数据集，以提高模型的泛化能力。
- **模型优化：**尝试更复杂的神经网络结构，如深层卷积神经网络、长短时记忆网络（LSTM）或基于 Transformer 的模型，以提高识别精度。
- **多语言支持：**扩展系统以支持多种语言的语音识别，提高其在不同语言环境下的适用性。

总体而言，本报告为连续数字串语音识别系统的设计和实现提供了一种有效的方法，并通过实验验证了其可行性，为后续的研究和开发工作奠定了基础。