实验一　自动识别技术——语音识别系统的设计与实现（5）

1. 语音转换为文本

前面的实验中，通过机器学习实现了简单短语的识别，但是，这种方式只能对整句或整个短语匹配才行，不够灵活，本实验将介绍实际的语音识别过程，并实现语音到文本的转换。

首先，了解本实现需要掌握的理论基础和技术原理。

1.1 音素（phoneme）

从声学性质来看，音素是从音质角度划分出来的最小语音单位。从生理性质来看，一个发音动作形成一个音素。例如〔ma〕包含〔m〕、〔a〕两个发音动作，是两个音素。相同发音动作发出的音就是同一音素，不同发音动作发出的音就是不同音素。

在语音识别过程中，对语音分帧并提取语音特征后，接下来就是要把这些帧组合成音素。为此，需要首先进行单音素的模型训练。

1.2 状态

但是，语音是一个连续的音频流，一个音素的前后音素的发音对本音素也是有影响的，即上下文是对音频波形影响较大的因素，这种叫做协同发音。所以我们需要根据上下文来辨识音素。

因此，将音素划分为三个状态：入音（音素的第一部分与在它之前的因素存在关联）、持续音（音素的中间部分是稳定的）、出音（音素的最后一部分与下一个音素存在关联）

所以，实际的语音识别过程为：

(1). 对帧进行特征提取，然后识别成状态

(2). 把若干状态（一般3个状态）组合成音素

(3). 把若干音素组成成字母文本

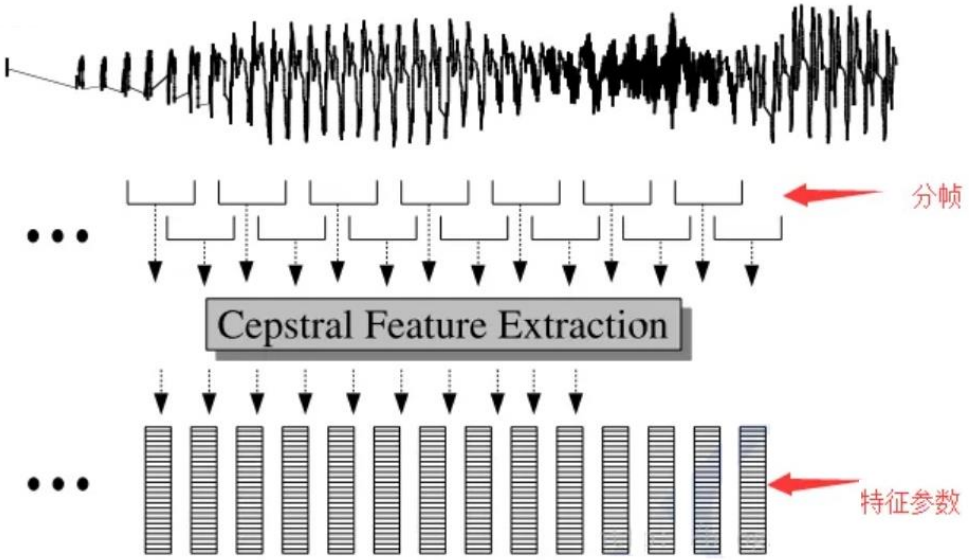
(4). 将字母文本转换为文字文本（中文识别）

1.3 MFCC-GMM-HMM 声学模型（语音识别过程）

MFCC，在前面实验中提到过，我们把一段语音信号分解为很多很多的小段语音片段（语音帧），例如每帧长度为25ms，相邻两帧有10ms重叠，也就是说帧长25ms，帧移10ms。

然后，我们对每一帧做信号分析，即特征提取，常见的特征参数有：MFCC、PLP。MFCC（Mel频率倒谱系数）算法将语音信号分帧后，对每一帧进行傅里叶变换，然后将频率

轴上的能量值转换为Mel频率轴上的能量值，并计算其倒谱系数。最后，将每一帧的倒谱系数串联起来，得到该帧的MFCC特征向量。经过MFCC特征提取后，压缩了特征参数。



接下来，考虑如何把一系列特征参数序列转化为一段话，也就是声学模型（GMM-HMMs）。我们知道，一段话是由一串文字序列组成，一个文字由一串音素组成一个音素对应一个HMM(隐马尔科夫模型)，同时通常一个HMM由三个状态（state）组成。那么，一个特征参数序列，识别的过程，就是解决怎么把每个特征参数识别为一个状态，再由状态到音素，音素到单词，单词到单词序列（一段话）。其中特征参数到状态，由GMMs（高斯混合模型）解决；三个状态到一个音素，由HMM解决；音素到单词，由词典解决；单词到单词序列，由语言模型解决。

GMM（Gaussian mixture model）；聚类算法。对mfcc特征进行聚类处理，获得M个类（类中心）。

GMM广泛用于聚类、数据分类、异常检测、图像分割和图像生成等任务。GMM通常使用期望最大算法 (Expectation Maximization,简称EM)进行参数估计，这是一种有效的优化算法，能够有效地找到模型参数的最大似然估计。基本思想是通过模型来计算数据的期望值，并不断更新各个参数，使得期望值最大化，通过迭代的方法直到两次迭代中参数变化十分微小为止。

HMM(隐马尔科夫模型)是比较经典的机器学习模型，在语音识别，自然语言处理、分词等序列标注领域广泛使用。它能对时序信息进行建模，是一种利用已知的观测序列来推断未知变量序列的模型。

好比计算机上的输入法功能，我正在计算机前打字，在键盘上敲出来的一系列字符就是观测序列，而我实际想输入的文字则是隐藏序列，输入法的任务就是通过我输入的一系列字符尽可能猜测我想要写的内容，并把最可能得词语放在最前面让我选择。

针对语音识别，假设语音有1000帧，每帧对应1个状态，每3个状态组合成一个音素，那么会组成300个音素，但因每帧很短，1000帧不可能有这么多音素。实际上，大多数相邻帧的状态应该是相同的。我们可以用HMM来预测最可能的隐性音素序列。

GMM-HMM是目前主流的声学模型，可以将声音转换成音素。但音素转换成文本还涉及一个语言模型。语言模型是使用大量文本训练出来的，对于语音识别来说非常重要，如果不使用语言模型，识别出来的结果基本也是看不懂的乱码。

以上是实际的语音识别大致过程，需要很大的网络以及强大的模型支持，所以目前大部分都是以云服务的形式进行语音识别。

1.4 常见的语音识别 ASR 工具

语音识别 ASR (Automatic Speech Recognition)实现过程中涉及到：噪声抑制、声学模型、语言模型和置信度评估等。常见的开源免费的语音识别 ASR 工具有：

1) Whisper

Whisper 是 OpenAI 在 22 年 9 月开源的一个语音识别系统，它使用从网络上收集的 68 万小时（98 种语言）、多任务监督数据进行训练。除了可以用于语音识别，Whisper 还能实现多种语言的转录，以及将这些语言翻译成英语。

2) FunASR

FunASR 是一个基础的语音识别工具包，提供了多种功能，包括语音识别（ASR）、语言模型、说话人验证等。

3) Tensorflow ASR

Tensorflow ASR 是一个使用 Tensorflow 2.0 作为深度学习框架来实现各种语音处理的语音转文本开源引擎。

2. 实验内容及要求

1) （选做）实验内容：为了帮助更好的理解和掌握GMM算法，请利用GMM算法对大家熟悉的鸢尾花数据集进行训练，并分别输出GMM分类结果和原始分类结果，以及模型score，分析GMM算法分类效果是否理想，如果不理想，请尝试修改程序，使得GMM的分类结果与原始分类结果基本一致，如下图

```
GMM分类结果：
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1
 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
原始分类结果：
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
GMM score:0.9666666666666667
```


上传要求

(1). GMM算法运行效果图

(2). 源代码文件

2) （必做）实验内容：使用一种开源的ASR工具，识别本地音频文件的内容，以文本形式返回识别结果（语音转文本）。

上传要求

(1). 语音转文本的运行效果图

(2). 源代码文件