《数字信号处理》语音识别系列实验

引言：

语言是人类最重要的交流工具，自动语音识别技术起源于 20 世纪 50 年代，最早的商用系统是 IBM 在 90 年代推出的 ViaVoice。经过半个多世纪的发展，语音识别技术目前已日趋成熟并成功应用到人们的日常生活之中，如苹果手机的 Siri 体验、科大讯飞的迅速崛起等。

语音是一种典型的、易于获取的一维时序信号，语音识别技术也是数字信号处理课程绝佳的实践途径。时间序列分析、快速傅里叶变换、滤波器设计等多项数字信号处理的教学内容在语音识别核心技术中均占有重要地位。本系列实验即面向语音识别基本任务，由浅入深，循序渐进地设计完善语音识别系统，包括时域法、频域法、说话人识别三个具体实验。

实验 1 基于时域分析技术的语音识别

实验目的：熟悉语音数据的基本形式及特点，理解并应用离散时间信号的基本分析、处理方法，理解语音识别技术的概貌，为后续实验打好基础。

实验原理及要点：

1．语音信号的采集：

采集“0”、“1”、…、“9”这 10 个语音的 wav 文件，每个类别应采集 10 组以上的样本。可以通过 Windows 的录音机等应用软件来实现，也可以借助语音处理的 API 函数，通过编程的方式来实现。

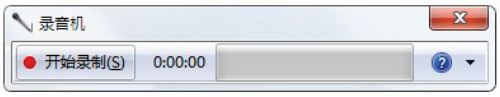


图 1. Windows 的录音机。

2．语音信号格式的理解：

通过互联网调研 wav 文件的具体格式，找到并理解其中与本任务密切相关的字段，如采样率等，能够编程实现对其中语音数据字段的读取功能。

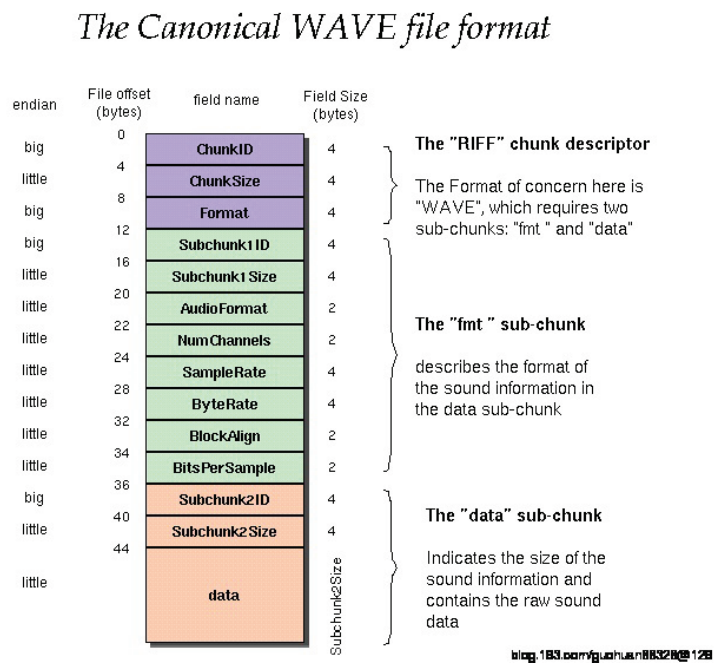


图 2. WAV 文件格式。

	字节数	具体内容	
ID	4 Bytes	'fmt '	
Size	4 Bytes	数值为16或18，18则最后又附加信息	
FormatTag	2 Bytes	编码方式，一般为0x0001	
Channels	2 Bytes	声道数目，1--单声道；2--双声道	
SamplesPerSec	4 Bytes	采样频率	
AvgBytesPerSec	4 Bytes	每秒所需字节数	===> WAVE_FORMAT
BlockAlign	2 Bytes	数据块对齐单位(每个采样需要的字节数)	
BitsPerSample	2 Bytes	每个采样需要的bit数	
	2 Bytes	附加信息（可选，通过Size来判断有无）	

图3 Format Chunk

3．语音信号的预处理：

对语音原始数据实现端点检测等基本的预处理任务，为后续的时域分析做好准备。端点检测的含义为将数据的实际发声部分从静音及背景噪声中分割出来，如图 1 所示。后续计算将仅针对分割出的部分进行。该部分可以采用交互式的手工方式来实现，也可编程自动化地实现。

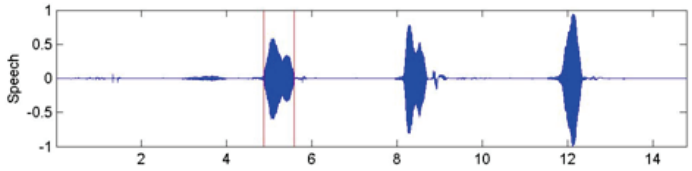


图 1. 语音数据的端点检测。红线部分即为对三个孤立语音中的第一个实现了端点检测。

4．时域分析：

基于已经提取的语音数据数组，对其时域特性进行分析和计算，可以计算其短时能量、过零率或其他你认为对本任务有益的数字特征。

其中短时能量的计算公式为：

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

这里 N 为帧长，E_n 表示第 n 帧语音信号 $x_n(m)$ 的短时能量。

第 n 帧语音信号过零率的计算公式为：

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

其具体计算步骤如下：

a） 首先对信号进行去直流化；

b） 然后按照时间顺序统计采样点数值符号变号的次数；

c） 将上述计数出的次数针对序列时长进行归一化操作，即得到过零率。

过零率实质上是信号频谱分布在时域的一种最简单的体现，即高频分量丰富的信号其过零率也一般较高。

5．语音识别分类器的实现：

针对上述提取完成的语音特征向量，选取合适的分类器算法来实现自动语音判别。可供选择的分类器包括 Naïve Bayesian、Fisher 线性判别、决策树、支撑向量机、最近邻分类器等。分类器的选取应充分说明理由，并在下述实验中通过对比来支撑自己的观点。

6．实验对比及量化分析：

通过一定数量的实验结果，分析上述各个环节中算法的性能，并通过对比不同方法，验证所选用方法的优势。对于语音识别的精度应通过正确率、误纳率等各种指标进行统计分析与对比。实验结果应通过图、表、文字等多种方式进行综合呈现。

实验内容及要求：

1. 实现时域法语音识别的基本过程，允许对其中的部分环节采用手工交互式的方式来实现，但对于时域分析计算、分类器实现等核心模块应编程实现。
2. 编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab 等。
3. 本实验对于界面编程不做具体要求。
4. 实现平台不限，Windows、Linux、或 Android 均可。
5. 以小组为单位，完成一份实验报告，报告应遵循学术论文的一般格式和规范。

实验 2 基于频域分析技术的语音识别

实验目的：熟悉语音数据的基本形式及特点，理解并应用离散时间信号的基本分析、处理方法，理解语音识别技术的概貌，为后续实验打好基础。

实验原理要点：

1．语音信号的频谱计算：

傅里叶变换是频谱计算的主要途径，而 FFT 则是工程上实现傅里叶变换的利器。本实验就要用到 FFT，注意帧长应设置为 2 的整数次幂，以利于实现按时间抽取或按频率抽取的蝶形算法。

2．梅尔（Mel）频率谱的计算：

梅尔(Mel)频率谱是在已知信号频谱的基础上，基于人类听觉系统的感知特性，设计出的一种频谱分组方式。通过计算 Mel 频谱，将得到比原始傅里叶频谱更加具有区分性的频域紧凑表达，从而有利于精确地实现识别任务。

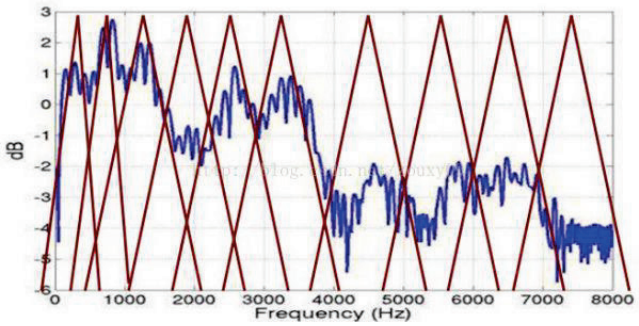


图 2. 梅尔频率倒谱的基本概念示意图。

3．语音信号的预加重：

由于语音的高频分量对于识别具有特别的意义，然而高频分量又通常能量较弱，因此应对原始语音信号首先进行预加重滤波处理，再进行后续的频谱计算。这就涉

及到数字滤波器的类型及参数选择。

4．DTW 技术的应用：

由于每一个孤立语音信号的时长一定各不相同，其计算得到的频谱特征向量长度也将各不相同。然而，对于一般的模式识别系统而言，要求待比对的特征向量应具有相同的长度。Dynamic Time Warping （DTW）技术则基于动态规划的思想，可以实现不等长特征向量的距离计算，因此在语音识别中得到了广泛应用。

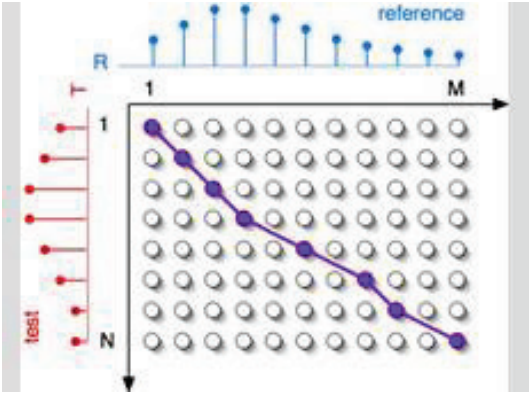


图 3. DTW 算法的基本思想是基于最短路径的搜索。

5．实验对比及量化分析：

通过一定数量的实验结果，分析上述各个环节中算法的性能，并通过对比不同方法，验证所选用方法的优势。对于语音识别的精度应通过正确率、误纳率等各种指标进行统计分析与对比，体会频域方法相比于时域方法在语音识别性能上的巨大提升效果。实验结果应通过图、表、文字等多种方式进行综合呈现。

实验内容及要求：

1. 仍旧面向“0”-“9”这 10 个孤立语音的识别任务，实现频域法语音识别。各个功能模

块均应采用编程来实现，包含必要的界面，能够自动地完成语音识别的完整过程。

2. 编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab、Qt 等，或是多种语言的混合编程。

3. 实现平台不限，Windows、Linux、Android、IOS 均可。

4. 以小组为单位，完成一份实验报告，报告应遵循学术论文的一般格式和规范。

实验 3 独立于内容的说话人识别

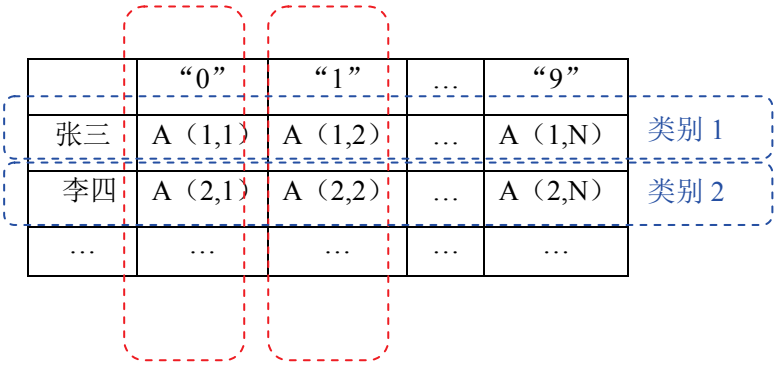
实验目的：在上述实验的基础上，实现一个更具挑战性的任务。旨在锻炼文献调研、开拓性思考、问题解决的能力，提升科研素养，并深化对信号处理、模式识别技术的理解和掌握。

实验原理要点：

1．与实验 1&2 的区别：

任何语音信号（signal）都具有语言内容（content）和说话人（speaker）两个基本属性。前述两个实验都是以估计内容为目标的，而本实验将面向估计说话人。更进一步，前述实验估计的是封闭的信号集（如 0 至 9），而本实验并不限制说话的内容，转而估计发音的主体（人），即独立于内容。

2．理解的误区：



王五	A（M,1）	A（M,2）	…	A（M,N）

类别 1　　类别 2

图 4. 对本实验任务一种肤浅的理解

上图示意了对本任务一种错误的理解。假设 M 个说话人都来发音“0”至“9”共 N 个数字，则其全部语音数据可以构成如上一个 M*N 的 A 矩阵。将这些信号都完成特征提取（如 FFT+Mel）后，对于实验 1&2，则其本质可以理解为红色字体类别 1、类别 2、…、类别 N 的分类问题，可以通过 DTW、kNN 等距离度量算法实现分类目的。这里的不同说话人也可用来表达同一说话人的不同次数据采集结果。

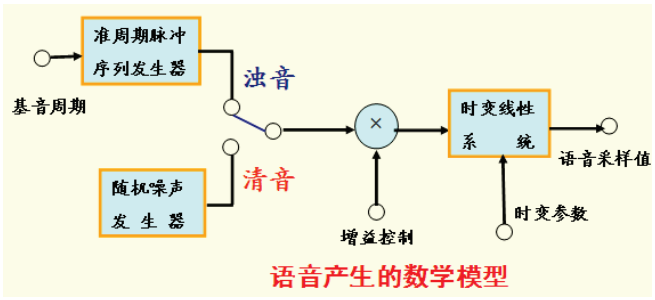
那么，实验 3 是否可以也采用完全相同的技术路线，转而理解为对蓝色字体类别 1、类别 2、…、类别 M 的分类问题呢？

答案显然是否定的，因为这样虽然也实现了说话人识别，但却依赖于闭集的语音信号集，即需要限定说话人的“0”至“9”这几个有限的语音信号，即可以称为“文本相关的说话人识别”。这与本实验所要求的“独立于说话内容”是有区别的。

3．实现思路的提示：

实际上人类具有非常完美的说话人识别能力。想想我们每次听到电话对端那熟悉的语音，无论对方说的是什么内容，是不是说话人的形象早已浮现在你的脑海之中！本实验就是要模仿这种能力，这将是一个比实验 1&2 更具挑战性的内容。

相信经过前述实验的积累，同学们已经掌握了语音信号特征提取的基本手段，这些特征在我们本实验中仍将起着重要的作用。但本实验的重心将转向这些特征向量所构成特征空间的建模、分析和分类/系统函数设计。



语音产生的数学模型

为了能够从任意语音信号中挖掘出能够代表说话人声门及声道排他性特点的微妙特征，请仔细观察上图中语音产生的数学模型，思考解决思路。提示以下三个可能的技术切入点：

a）大数据分析：利用模型训练阶段数据量的庞大，来尽可能覆盖在线识别阶段所能遇到的所有数据类别。但多类别分类器的实时性将是一个需要考虑的主要问题。

b）非线性分类器：语音内容（content）与说话主体（identity）相交织的特征空间将呈现出复杂的数据分布模式，这些数据间往往难以实现线性可分。而众多强有力的非线性分类技术（如核 SVM）是解决这一问题的利器。

c）线性预测分析技术：内容可以视为信号 x，一个人的声门声道可以视为系统 h，发出的语音则是输出 y。信号通过系统是利用卷积运算实现的，现在为了从大量的 x 和 y 中估计出系统函数，线性预测分析（LPC）、最小平方逆滤波、盲反卷积等都是可供选择的技术。

实验内容及要求：

1. 本实验为选做实验，仅供学有余力的同学思考和实践。因此本实验结果并不计入期末的总成绩内。
2. 可以将说话人局限为本小组内的 3-5 人（即小类别个数），但不应限制说话的内容，例如可以通过说话人一段随机选取的普通新闻短稿朗读，判断出说话人身份。
3. 请通过文献调研与小组讨论的方式了解和掌握面向本任务的基本实现算法，并编程实现

其基本功能。界面可以沿用实验 1&2 的类似风格，并非本实验的重点。

4. 编程语言不做要求，实现平台不限。
5. 对于实现了本实验要求的小组，请完成一份实验报告，并遵循学术论文的一般格式和规范。