# 数字信号处理

# 实验指导书

（2020 版）

姓名：_____

学号：_____

本实验课程是本科生《数字信号处理》课程的专题实验。

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，语音信号处理及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，包括语音信号采集与预处理、基于时域分析技术的语音识别、语音信号的频域特征分析、基于动态时间规整（DTW）的孤立字语音识别实验，共分为四个实验。

本实验课程通过语音信号的采集处理及识别系统设计实验使学生巩固和加深数字信号处理的理论知识，采用这种研究型实验、并辅之以课堂测验及讨论等教学手段，进一步加强学生独立分析问题、解决问题的能力，培养综合设计及创新能力，培养学生实事求是、严肃认真的科学作风和良好的实验习惯，为今后的工作打下良好的基础。

# 实验 1 基于时域分析技术的语音识别

## 一、实验目的：

1、理解语音信号的采样与量化对语音信号质量的影响，了解语音信号格式；

2、掌握并编程实现常用的窗函数和加窗分帧处理方法；

3、掌握语音短时时域分析的原理及时域参数计算方法；

4、掌握双门限法的端点检测法；

5、掌握并编程实现基于时域分析技术实现孤立字语音识别方法。

## 二、实验环境与实验器材

1、实验硬件设备列表

计算机、耳麦

2、实验软件配置列表

（1）编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab 等；

（2）实现平台不限，Windows、Linux 或 Android 均可；

（3）实验室计算机系统为 Windows、软件为 Matlab。

## 三、实验原理

### 1、语音信号的采集与量化：

为了将原始模拟信号变为数字信号(A/D 转换)，必须经过采样与量化两个步骤，从而得到时间和幅度上均为离散的数字信号。采样时必须要满足奈奎斯特定理，即采样率$f_s$必须以高于受测信号的最高频率两倍以上的速度进行取样，才能正确地重建信号。

**2、语音信号格式的理解：**WAV 文件格式如图 1.2、图 1.3

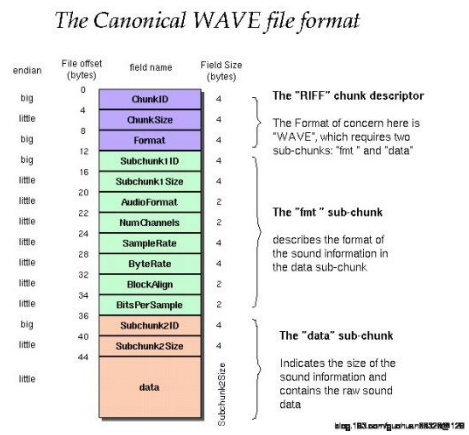



图 1.2　WAV 文件格式

| | 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 1.3　WAV 文件字段

### 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 1.4 所示。

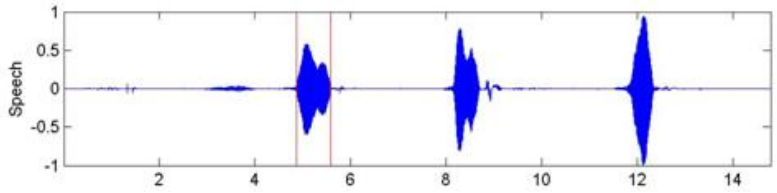



图 1.4　语音数据的端点检测

**4、语音分帧及加窗：**

语音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，所以它是一个非平稳态过程，不能用处理平稳信号的数字处理技术对其进行处理。但是，由于不同的语音是由人的口腔肌肉运动构成声道某种形状而产生的响应，而这种口腔肌肉运动相对于语音频率来说是非常缓慢的，所以从另一个方面看，虽然语音信号具有时变特性，但是在一个短时间范围内（一般认为在 10~30 ms 的时间内），其特性基本保持不变即相对稳定，因而可以将其看做是一个准稳态过程，即语音信号具有短时平稳性。

任何语音信号的分析和处理必须进行“短时分析”，将语音信号分为一段一段来分析其特征参数，其中每一段称为一“帧”，帧长一般取 10~30 ms。这样，对于整体的语音信号来讲，分析出的是由每一帧特征参数组成的时间序列。

分帧是用可移动的有限长度窗口进行加权的方法来实现的，就是用一定的窗函数 $w(n)$ 来乘语音信号 $s(n)$，从而形成加窗语音信号 $s_w(n)=s(n)*w(n)$。窗函数 $w(n)$ 的选择（形状和长度），对于短时分析参数的特性影响很大。在语音信号数字处理中常用的窗函数有三种：

[1]　矩形窗

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

[2]　汉明窗

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

[3]　海宁窗

$$w(n)=\begin{cases}0.5(1-\cos(2\pi n/(N-1))), & 0\le n\le N-1\\ 0, & \text{其他}\end{cases}$$

**5、短时时域特性分析**

语音信号的时域分析就是分析和提取语音信号的时域参数。时域分析通常用于最基本

的参数分析及应用，如语音的分割、预处理、分类等。语音信号的时域参数有多种，本实验着重掌握短时能量、短时平均幅度及短时过零率。

[1]、 短时能量与短时平均幅度

设第 $n$ 帧语音信号 $x_n(m)$ 的短时能量用 $E_n$ 表示，则其计算公式如下：

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

$E_n$ 是一个度量语音信号幅度值变化的函数，但它有一个缺陷，即它对高电平非常敏感（因为它计算时用的是信号的平方）。为此，可采用另一个度量语音信号幅度值变化的函数，即短时平均幅度函数 $M_n$，它的定义为：

$$M_n = \sum_{m=0}^{N-1} |x(m)|$$

$M_n$ 也是一帧语音信号能量大小的表征，它与 $E_n$ 的区别在于计算时小取样值和大取样值不会因取平方而造成较大差异，在某些应用领域中会带来一些好处。

[2]、 短时过零率

短时过零率表示一帧语音中语音信号波形穿过横轴（零电平）的次数。对于连续语音信号，过零即意味着时域波形通过时间轴；而对于离散信号，如果相邻的取样值改变符号则称为过零。因此，过零率就是样本改变符号的次数。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。

设第 $n$ 帧语音信号 $x_n(m)$ 的短时过零率用 $Z_n$ 表示，则其计算公式如下：

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

式中，$\mathrm{sgn}[.]$ 是符号函数，即 $\mathrm{sgn}[x] = \begin{cases} 1, & (x \geq 0) \\ -1, & (x < 0) \end{cases}$

**6、双门限法的端点检测**

根据语音的统计特性，可以把语音段分为清音、浊音以及静音（包括背景噪声）三种。语音端点检测本质上是根据语音和噪声的相同参数所表现出的不同特征来进行区分。在双门限法中，短时能量可以较好地区分出浊音和静音。对于清音，由于其能量较小，在短时能量检测中会因为低于能量门限而被误判为静音；短时过零率则可以从语音中区分出静音和清音。将这两种检测结合起来，就可以检测出语音段（清音和浊音）及静音段。在基于短时能量和过零率的双门限端点检测算法中首先为短时能量和过零率分别确定两个门限，一个为较低的门限，对信号的变化比较敏感，另一个是较高的门限。当低门限被超过时，很有可

能是由于很小的噪声所引起的，未必是语音的开始，到高门限被超过并且在接下来的时间段内一直超过低门限时，则意味着语音信号的开始。

## 四、实验步骤及要求

### 1、语音信号采集与 WAV 文件格式理解

本实验采集“0”、“1”、…、“9”这 10 个语音的 wav 文件。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。在 MATLAB 环境中语音信号的采集可使用 audiorecorder 函数录制。

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。

### 2、语音数据集建立

利用语音信号频谱分量分布的特征，可以采用 8kHz 的采样率对语音信号进行采样，得到离散的语音信号。并对采集到的语音信号进行分析。

（1）每人采集 20 组以上语音（每组包含 0”-“9”10 个孤立语音）

（2）对语音进行预处理、截取和存储

（3）以班级为单位，构建包含 300 组以上的语音数据集，按二八原则分为测试集和训练集

### 3、采样与量化

改变采样周期和量化补偿，分析不同参数对语音质量的影响，并编程实现用图和数据形式分析采样率及量化精度对语音信号的影响。存储 WAV 文件时，可以分别以采样频率、2 倍的采样频率和 1/2 采样频率进行存储。量化精度可设置为 16 位、8 位等。

### 4、时域特征向量提取

（1）**分帧及加窗**

根据分帧思想，利用矩形窗、汉明窗、海宁窗分别对语音信号进行处理并对比三种加窗函数。分帧示意图如图 1.5 所示。一般每秒的帧数约为 33~100 帧，视实际情况而定。分帧虽然可以采用连续分段的方法，但一般要采用如图所示的交叠分段的方法，这是为了帧与帧之间平滑过渡，保持其连续性。前一帧和后一帧的交叠部分称为帧移。帧移和帧长的比值一般取为 0~1/2。

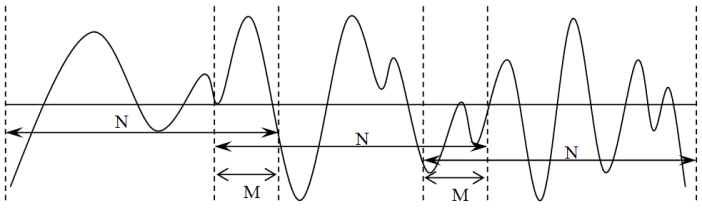


图 1.5　语音分帧示意图

**（2）双门限法的端点检测**

利用语音信号时域特性实现基于双门限法的语音端点检测，具体步骤如下（参见图 1.6）：

[1]、 计算信号的短时能量和短时平均过零率。

[2]、 根据语音能量的轮廓选取一个较高的门限$T_2$，语音信号的能量包络大部门都在此门限之上，这样可以进行一次初判。语音起止点位于该门限与短时能量包络交点$N_3$和$N_4$所对应的时间间隔之外。

[3]、 根据背景噪声的能量确定一个较低的门限$T_1$，并从初判起点（$N_3$）往左，从初判终点（$N_4$）往右搜索，分别找到第一次与门限$T_1$相交的两个点$N_2$和$N_5$，于是$N_2N_5$段就是用双门限法所判定的语音段。

[4]、 以短时平均过零率为准，从$N_2$点往左和$N_5$点往右搜索，找到短时平均过零率低于某阈值$T_3$的两点$N_1$和$N_6$，这便是语音段的起止点。

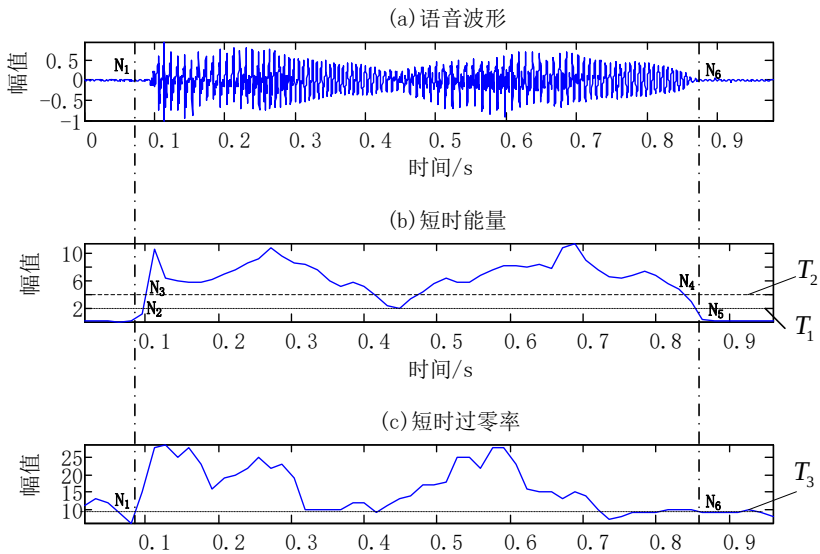



图 1.6 双门限法端点检测示意图

**（2）特征向量提取**

对去清音、浊音及静音段的语音，提取短时能量、短时平均幅度及短时过零率等时域特征参数，并构建语音信号的时域特征向量。

**5、基于时域分析技术孤立字语音识别：**

针对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。**（K 近邻分类算法、支持向量机算法见附录）**

**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检率、误检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。

## 四、实验报告要求

1、以小组（3-5 人）为单位，按所编写程序完成一份实验报告；

2、详细描述实验过程（实验步骤、现象描述、实验记录等）；

3、详细分析实验数据并对实验结果进行论述；

4、鼓励对实验提出进一步思考及建议的方向；

5、实验报告应遵循学术论文的一般格式和规范。

## 五、实验考核与成绩评价标准

1、总成绩=实验出勤（20%）+现场表现（60%）+实验报告（20%）

# 实验二 语音信号的频域特征分析

## 一、实验目的

1、掌握短时傅里叶变换原理，并编程实现短时傅里叶变换；

2、理解 Mel 滤波器的基本原理；

3、编程提取语音的 Mel 频率倒谱特征 MFCC。

## 二、实验环境与实验器材

1、实验硬件设备列表

计算机、耳麦

2、实验软件配置列表

（1）编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab 等；

（2）实现平台不限，Windows、Linux、或 Android 均可；

（3）实验室计算机系统为 Windows、软件为 Matlab。

## 三、实验原理

### 1、短时傅里叶变换

语音信号是一种典型的非平稳信号，但是其非平稳性是由发音器官的物理运动过程而产生的，此过程与声波振动的速度相对比较缓慢，可以假定在 10~30ms 这样的短时间内是平稳的。傅里叶分析是线性系统和平稳信号稳态特性的强有力手段，而短时傅里叶分析，是在短时平稳的假定下，用稳态分析方法处理非平稳信号的一种方法。

设语音信号时域信号为 $x(l)$，加窗分帧处理后得到的第 $n$ 帧语音信号为 $x_n(m)$，则 $x_n(m)$ 满足下式：

$$x_n(m) = w(m)x(n+m) \quad 0 \le m \le N-1$$

设离散时域采样信号为 $x(n)$，$n = 0,1,...,N-1$，其中 $n$ 为时域采样点序号，$N-1$ 是信号长度。然后对信号进行分帧处理，则 $x(n)$ 表示为 $x_n(m), m = 0,1,...,N-1$，其中 $n$ 是帧序号，$m$ 是帧同步的时间序号。信号 $x(n)$ 的短时傅里叶变换为：

$$X_n(e^{j\omega}) = \sum_{m=0}^{N-1} x_n(m)e^{-j\omega m}$$

定义角频率 $w = 2\pi k / N$，则得离散的短时傅里叶变换（DFT），它实际上是 $X_n(e^{jw})$ 在频域的取样，如下所示：

$$X_n(e^{j\frac{2\pi k}{N}}) = X_n(k) = \sum_{m=0}^{N-1} x_n(m) e^{-j\frac{2\pi km}{N}} \quad (0 \le k \le N-1)$$

在语音信号数字处理中，都是采用 $x_n(m)$ 的离散傅里叶变换 $X_n(k)$ 来替代 $X_n(e^{jw})$，并且可以用高效的快速傅里叶变换（FFT）算法完成由 $x_n(m)$ 至 $X_n(k)$ 的转换。当然，这是窗长 N 必须是 2 的倍数 $2^L$（$L$ 是整数），以便利于实现按时间抽取或按频率抽取的蝶形运算。

**2、Mel 频率倒谱系数**

与时域特征相比，频域特征更能辨识语音。

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。Mel 频率尺度的值大体上对应于实际频率的对数分布关系，与实际频率可用下式近似表示：

$$Mel(f) = 2595\lg(1 + f/700)$$

式中，$f$ 为频率，单位为 Hz。

语音频率可以划分成一系列三角形的滤波器序列，即 Mel 滤波器组，如图 2.1 所示。

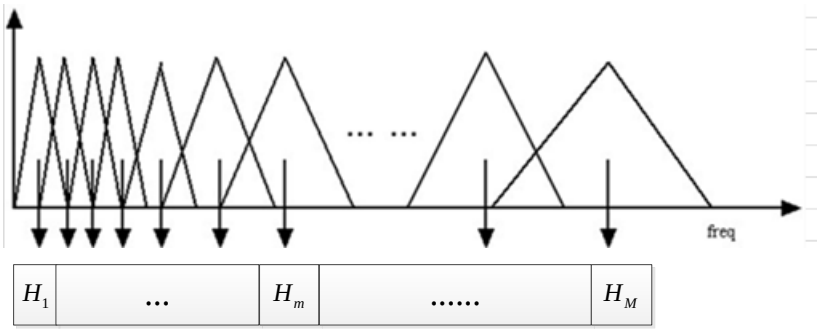



图 2.1 Mel 频率尺度滤波器组

设划分的带通滤波器为 $H_m(k)$，$0 \le m < M$，$M$ 为滤波器的个数。每个滤波器具有三角形滤波特性，其中心频率为 $f(m)$，在 Mel 频率范围内，这些滤波器是等带宽的（如图 2.2）。

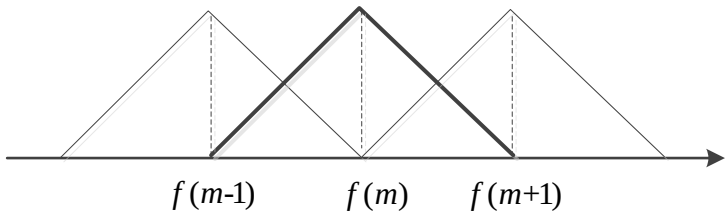


图 2.2 相邻三角滤波器之间的关系

每个带通滤波器的传递函数为：

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)\le k\le f(m+1)\\ 0 & k>f(m+1)\end{cases}$$

其中，$\sum_{m}^{M-1}H_m(k)=1$。

Mel 滤波器的中心频率 $f(m)$ 定义为：

$$f(m)=\frac{N}{f_s}F_{Mel}^{-1}(F_{Mel}(f_l)+m\frac{F_{Mel}(f_h)-F_{Mel}(f_l)}{M+1})$$

其中，$f_h$ 和 $f_l$ 分别是滤波器组的最高频率和最低频率，$f_s$ 为采样频率，单位为 $Hz$。

$M$ 是滤波器组的数目，$N$ 为 FFT 变换的点数，$F_{Mel}^{-1}(b)=700(e^{\frac{b}{1125}}-1)$。

## 四、实验步骤及要求

**1、根据短时傅里叶变换的原理，编写其函数，自行实现 FFT 过程。**

**2、根据 Mel 频率倒谱系数原理，提取语音特征参数 MFCC。**

具体流程参见图 2.3。

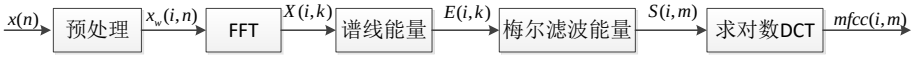



图 2.3 MFCC 参数提取基本流程

（1）预处理包括预加重、分帧、加窗

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。预加重的滤波器常用 $H(z)=1-az^{-1}$，式中，$a$ 是一个常数，如 0.97。

按照实验一相关步骤对语音信号进行分帧、加窗处理。

（2）快速傅里叶变换

对每一帧信号信息进行 FFT 变换，从时域数据转变为频域数据。$X(i,k)=FFT(x_i(m))$

（3）计算谱线能量

计算每一帧 FFT 后的数据谱线能量： $E(i,k)=[X_i(k)]^2$

（4）计算 Mel 滤波器能量

把求出的每帧谱线能量通过 Mel 滤波器，并计算在该 Mel 滤波器中的能量。在频域中相当于把每帧的能量谱 $E(i,k)$（其中 $i$ 表示第 $i$ 帧，$k$ 表示频域中的第 $k$ 条谱线）与 Mel 滤波器的频域响应 $H_m(k)$ 相乘并相加：

$$S(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),0\le m<M$$

（5）经离散余弦变换（DCT）得到 MFCC 系数：

$$mfcc(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\ln[S(i,m)]\cos[\frac{\pi n(2m-1)}{2M}]$$

式中，$S(i,m)$ 是（4）求出的 Mel 滤波器能量；$m$ 是指第 $m$ 个 Mel 滤波器（共有 M 个）；$i$ 是指第 $i$ 帧， $n$ 是 MFCC 系数阶数。提取频谱包络：选取低阶 MFCC 系数，n 通常取 12-16。

语音的 MFCC 频域特征示意图如图 2.4:

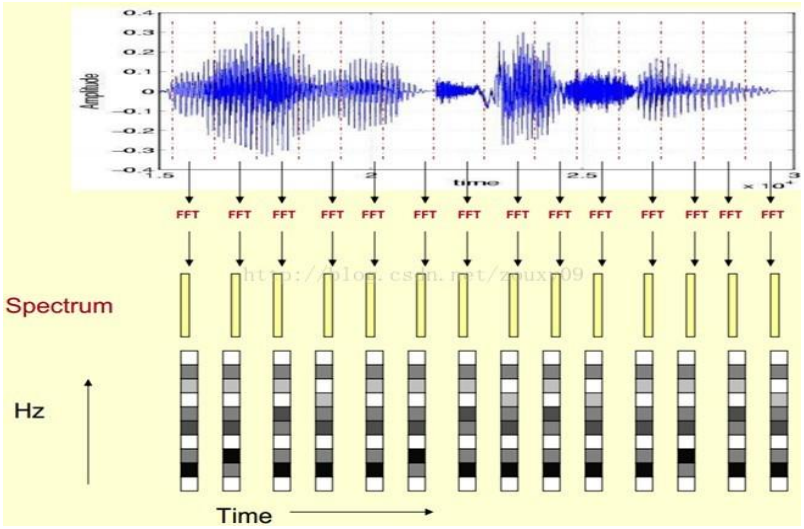



图 2.4 语音的 MFF 频域特征示意图

## 五、实验考核与成绩评价标准

1、总成绩=实验出勤（20%）+现场表现（60%）+实验报告（20%）

# 实验三 基于动态时间规整（DTW）的孤立字语音识别实验

## 一、实验目的

1、掌握语音识别的模板匹配法的原理和过程；

2、 掌握动态时间规整（DTW）技术；

3、 应用 MATLAB 实现基于 DTW 的 10 个阿拉伯数字的识别，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。

## 二、实验环境与实验器材

1、实验硬件设备列表

计算机、耳麦

2、实验软件配置列表

（1）编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab 等；

（2）实现平台不限，Windows、Linux 或 Android 均可；

（3）实验室计算机系统为 Windows、软件为 Matlab。

## 三、实验原理

### 1、模板匹配法语音识别系统构成

图 3.1 为利用模板匹配法进行语音识别的原理框图。在训练阶段，用户将词汇表中的每一个词一次说一遍，并且将其矢量特征时间序列作为模板存入模板库；在识别阶段，将输入语音的特征矢量时间序列依次与模板中的每一个模板进行相似度比较，将相似度最高者作为识别结果输出。

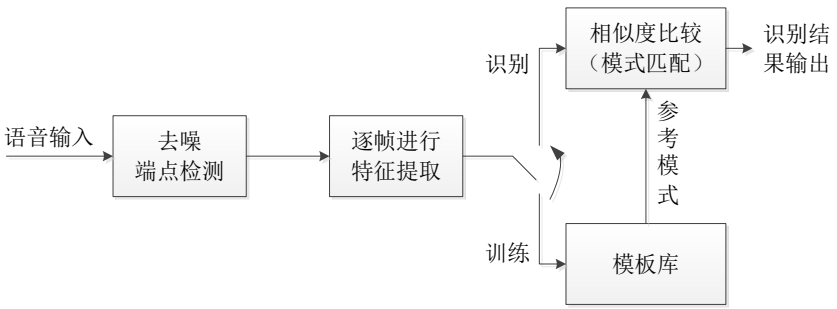



图 3.1 模板匹配法语音系统的原理框图

在特征提取阶段，本实验选用实验二的 Mel 频率倒谱系数（MFCC）作为识别特征；在识别阶段，实验选用动态时间规整（DTW）技术进行模式匹配。

### 2、动态时间规整（DTW）

由于每一个孤立语音信号的时长各不相同，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）技术是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。

动态时间规划是一个典型的最优化问题，它用满足一定条件的时间规整函数描述输入模板和参考模板之间的时间对应关系，求解两模板匹配时累计距离最小所对应的规整函数。假设词库中的某一参考模板的特征矢量列为 $a_1,...,a_m,...,a_M$ ，输入语音的特征矢量序列为 $b_1,...,b_n,...,b_N, M \neq N$ ，那么动态时间规整就是要找到时间规整函数 $m = T(n)$ 。该函数把输入模板的时间轴 $n$ 非线性映射到参考模板的时间轴 $m$ ，并且满足下式：

$$D = \min_{T(n)} \sum_{n=1}^{N} d[n, T(n)]$$

式中，$d[n, T(n)]$ 表示两帧矢量之间的距离；$D$ 是最佳时间路径下两个模板的距离测度。本实验中距离测度可采用欧式距离：

$$d(x, y) = \frac{1}{k}\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \text{ 。}$$

## 四、实验步骤与要求

### 1、DTW 算法的具体实现

DTW 算法的原理图如图 3.2 所示，把测试模板的各个帧号 $n = 1 \sim N$ 在一个二维直角坐标系中的横轴上标出，把参考模板的各帧 $m = 1 \sim M$ 在纵轴上标出，通过这些表示帧号的整数坐标画出一纵横线即可形成一个网络，网络中的每一个交叉点表示测试模式中某一帧与训练模式中某一帧的交汇。DTW 算法分两步进行，一是计算两个模式各帧之间的距离，即求出帧匹配距离矩阵，二是在帧匹配距离矩阵中找出一条最佳路径。搜索这条路径的过程可以描述如下：所搜从 $(1,1)$ 点出发，对于局部路径约束，点 $(i_n, i_m)$ 可到达的前一个格点可能是 $(i_{n-1}, i_m)$ ，$(i_{n-1}, i_{m-1})$ 和 $(i_n, i_{m-1})$ 。那么 $(i_n, i_m)$ 一定选择这三个距离中的最小者所对应的点作为其前续格点，这时此路径的累积距离为：

$$D(i_n, i_m) = d(T_i(i_n), R(i_m)) + \min\{D(i_{n-1}, i_m), D(i_{n-1}, i_{m-1}), D(i_n, i_{m-1})\}$$

这样从 $(1,1)$ 点 $(D(1,1) = 0)$ 出发搜索，反复递推，直到 $(M, N)$ 就可以得到最优路径，而且 $D(M, N)$ 就是最佳匹配路径所对应的匹配距离。在进行语音识别时，将测试模板与所有参考模板进行匹配，得到最小匹配距离 $D_{\min}(N, M)$ 所对应语音即为识别结果。

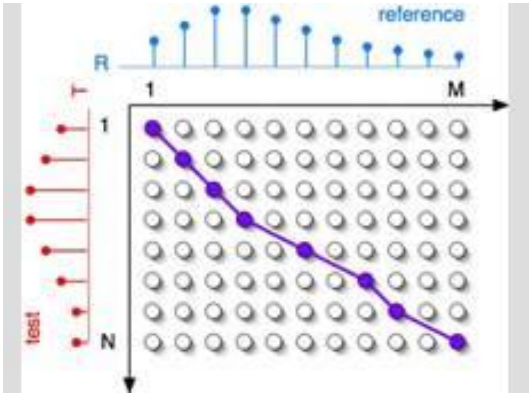



图 3.2 DTW 算法原理图

2、面向“0”-“9”这 10 个孤立语音的识别任务，利用实验二获取的 MFCC 系数，实现频域法语音识别。

## 五、实验报告要求

1、以小组（3-5 人）为单位，按所编写程序完成一份实验报告；

2、详细描述实验过程（实验步骤、现象描述、实验记录等）；

3、详细分析实验数据并对实验结果进行论述；

4、鼓励对实验提出进一步思考及建议的方向；

5、实验报告应遵循学术论文的一般格式和规范。

## 六、实验考核与成绩评价标准

1、总成绩=实验出勤（20%）+现场表现（60%）+实验报告（20%）

# 实验四 独立于内容的说话人识别

## 一、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。

1、实验硬件设备列表

计算机、耳麦

2、实验软件配置列表

（1）编程语言不做要求，可以是 C/C++/C#、java、Pascal、Python、Matlab 等；

（2）实现平台不限，Windows、Linux 或 Android 均可；

（3）实验室计算机系统为 Windows、软件为 Matlab。

## 二、实验原理要点

**1、与实验 1&3 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：**

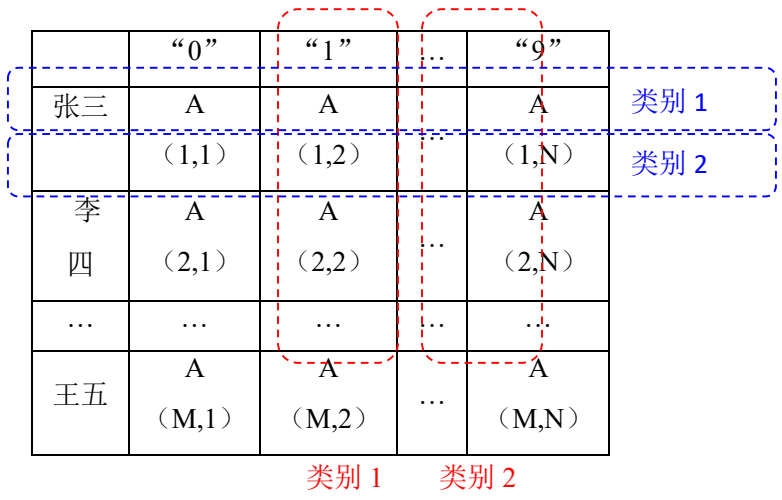



图 4.1 对本实验任务一种肤浅的理解

上图 4.1 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。

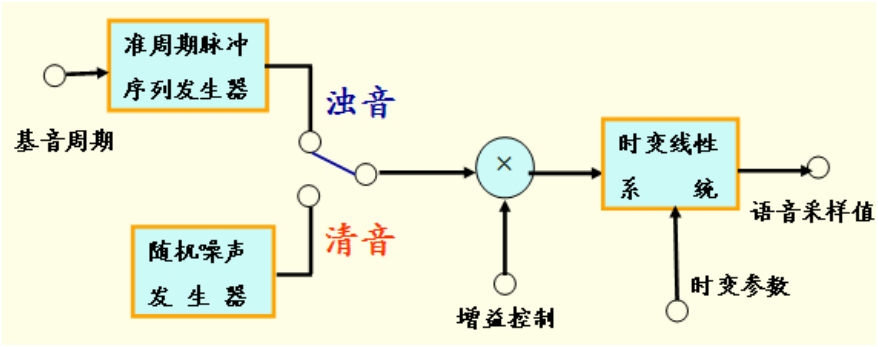



图 4.2 语音产生的熟悉模型

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 4.2 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：

（1）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。

（2）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。

（3）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。

## 三、 实验内容及要求：

1、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。

2、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&3 的类似风格，并非本实验的重点。

## 四、 实验考核与成绩评价标准

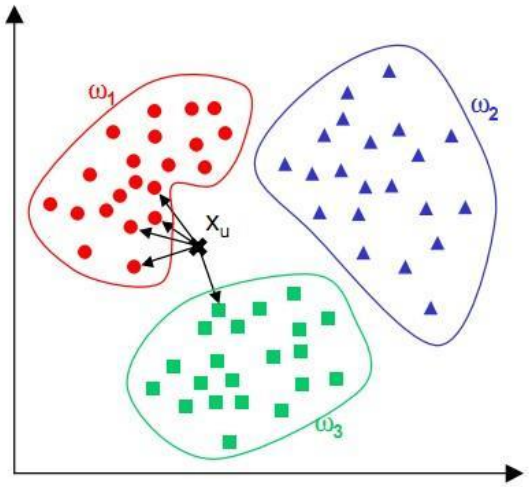
1、本实验为选做实验，仅供学有余力的同学思考和实践。因此本实验结果并不计入期末的总成绩内。

2、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。

**附录：模式识别基本方法**

**（1）k 近邻分类算法**

K 近邻（K-Nearest Neighbor，KNN）分类算法简单直观。其基本思想是：给定一个在特征空间中的待分类的样本，如果其附近的 k 个最邻近的样本中的大多数属于某一个类别，那么当前待分类的样本也属于这个类别。如附图 1.1 所示




附图 1.1 KNN 算法基本思想

KNN 算法大致可分为以下四步：

1）由特征提取函数提取训练样本的特征向量，构成训练样本的特征向量集合 $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, X_{\mathrm{n}}$ 。

2）设定算法中 K 的值。K 值的确定没有一个统一的方法（根据具体问题选取的 k 值可能有较大的差别）。一般先确定一个初始值，然后根据实验结果不断调试，最终达到最优。

3）利用特征向量提取函数待测样本的特征向量 X，并计算 X 与$\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, X_{\mathrm{n}}$中每一样本的欧式距离$D(X, X_l)$，$l = 1,2, \ldots, \mathrm{n}$。

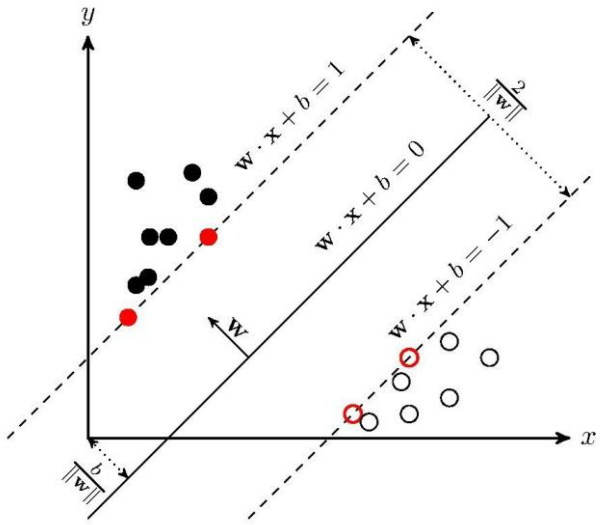
4）统计$D(X, X_l)$，$l = 1,2, \ldots, \mathrm{n}$中 K 个最近邻的类别信息，给出 X 的分类结果。

**（2）支持向量机算法**

支持向量机（SVM）算法是一种基于统计理论的学习方法，其基本思路就是要找到使测试样本的分类错误率达到最低的最佳超平面，也就是要找到一个分割平面，使得训练集中的训练样本距离该平面的距离尽量的远，该分割平面的两侧的空白区域最大。

对于两类问题进行分类时，存在多个超平面将两类样本分开，假设类别距超平面的距离为类别中所有样本距离超平面的最小值，则两个类别距所有超平面的距离最大的超平面称

为最优超平面。如下附图 2.2 所示，$w \cdot x + b = 0$即为最优超平面。




附图 2.2 最优分类超平面

如果样本是现象不可分的，则可引入松弛变量得到近似的线性超平面，或通过非线性映射算法实现低维输入空间线性不可分样本到高纬特征空间线性可分样本的映射，再进行分析。

对于 N 类问题进行分类，则需要对 SVM 进行组合。组合策略有“one-vs-one”和“one-vs-all”。“一对多”的思想是在该类样本和不属于该类样本之间构建一个超平面，假设总共有 K 个类别，则需构建 K 个分类器，每个分类器分别用第 i 类的样本作为正样本，其余的样本作为负样本。该方法的缺点是样本数目不对称，负样本比正样本要多得多，故分类器的训练中惩罚因子很难选择。“一对一”的方式是没两类样本间构造一个超平面，一共需要$k(k-1)/2$个分类器，最后识别样本时采用后验概率最大，从而选定待识别样本的类型，“一对一”的方法的缺点是训练的分类器比较多。