声纹识别

目前应用于声纹识别中的模式匹配识别技术主要有：矢量量化（VQ）、动态时间规整（DTW）及支持向量机（SVM）、高斯混合模型（GMM）和隐马尔可夫模型（HMM）及人工神经网络模型（ANN）。

作为一种与日常生活息息相关的密码指令身份确认系统，极有可能会在不同的终端设备上录制训练和测试的语音，且如若冒认者在目标人使用它时，用设备将其语音信息录制以待下次识别使用，本文所设计的声纹密码锁还需要解决不同类型的电脑、声卡以及手机麦克风所带来的不同信道的失配问题。

因此，如何在声纹识别的预处理阶段有效的去除复杂多变的噪声，提取到相对纯净的语音数据以获得更高的识别效率是本文声纹密码锁应用重要难题之一。

声纹常用特征提取算法 LPCC （线性预测倒谱系数）和 Mel 频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）特征提取进行研究。LPC 和 LPCC 的基本思想都是：语音信号当前样点的值可以用过去若干个样点的值的线性组合来逼近，通过使实际的值与线性预测的值的均方误差达到最小来确定一组唯一的参数作为语音的特征参数。它们的缺点在于所有频率上是线性逼近语音的，而实际上，人耳的听觉特性是一个非线性系统，它对不同的频率的灵敏度是不同的，基本上是对数关系。 Mel 倒谱参数(Mel. scaled Cepstrum Coefficiems)，即 MFCC，充分利用人耳感知不同频率语音的特殊性，将语音的频域从线性频标转换为 Mel 频标，符合人耳的听觉特性，提高了系统的识别性能。MFCC 和一般频率 f 的关系式如下(音调，是人耳对声音频率高低的感受，Mel 是音调的度量单位)：

$$\mathrm{mel(F)} = 2595 * \log_{10}\left(1 + \frac{\mathrm{f}}{700}\right)$$ 为对数关系，而非直线关系。

LPCC 的计算量较低，比较适合嵌入式语音识别、声纹识别方面的应用，而 MFCC 作为语音特征参数，更能体现语音的特点，符合人体听觉系统，因此本文语音特征提取方面选择 MFCC 特征参数。

贯穿语音分析全过程的是短时分析技术。语音信号特征随时间变化，是非平稳随机过程。但从另一方面看，虽然语音信号有时变特性，但短时间内其特性基本不变。这是因为人的肌肉运动具有惯性，从一个状态到另一个状态的转变不可能瞬时完成，而存在一个过程。在较短时间内语音信号的特征基本保持不变。基本采用短时平稳假设。因此，语音分析和处理建立在短时基础上，即对语音信号流进行分段处理，即分帧。

1　预处理

由于语音信号的时变性及非平稳性，且为便于计算，在进行端点检测前，需要对输入语音信号进行预加重、加窗和分帧等预处理过程。（根据取样定理，取样频率大于信号两倍带框时，取样过程不会丢失信息，且由取样信号可以精确地重构原信号）。

白噪声（white noise）是指功率谱密度在整个频域内是常数的噪声。若信号波形变化足够大或量化间隔足够小，可证明量化噪声符合具有下列特征的统计模型：（1）为平稳白噪声过程（2）量化噪声与输入信号不相关；（3）量化噪声在量化间隔内均匀分布，即具有等概率密度分布。量化阶梯足够小的话，信号幅度从一个取样值到相邻取样值的变化可能非常大，这样的量化噪声逼近与白噪声，与上述三个假设相吻合。

1. 语音信号的预加重，目的是**为了对语音的高频部分进行加重，去除口唇辐射的影响，增加语音的高频分辨率**。一般通过传递函数为

$$H(z)=1-\alpha z^{-1}$$

以$S_1(\mathrm{n})(n:0\ldots N-1)$表示时域信号，预增强公式为$\mathrm{S(n)}=S_1(\mathrm{n})-\mathrm{a}*\mathrm{S_1(n-1)}(0.9<a<1.0)$ 该过程可以达到在音框化阶段对静音数据的判断，因为静音数据的值是几乎不变的所以在做差分以后值会很小，接近于 0，而有声音的数据则会保留较大的值。

2. 对语音信号进行**加窗分帧处理可实现语音信号短时平稳性，一般会取33~100 帧/秒**。语音信号的分帧是通过加窗来实现的，在语音信号的处理中，窗函数通常包括矩形窗和汉明窗。

汉明窗：（假设音框化的信号（M 帧共 N 点）为 S(n)，n=0，1，…，N-1。那么乘上汉明窗后为: S′ (n) = S(n) ×W(n)）

汉明窗的主瓣宽度较宽，是矩形窗的一倍，但是汉明窗的旁瓣衰减较大，具有更平滑的低通特性，能够在较高的程度上反应短时信号的频率特性。另外，可防止吉布斯现象。吉布斯现象 Gibbs phenomenon（又叫吉布斯效应）： 将具有不连续点的周期函数(如矩形脉冲)进行傅立叶级数展开后，选取有限项进行合成。当选取的项数越多，在所合成的波形中出现的峰起越靠近原信号的不连续点。当选取的项数很大时，该峰起值趋于一个常数，大约等于总跳变值的 9%。这种现象称为吉布斯现象。

矩形窗的主瓣宽度小于汉明窗，具有较高的频谱分辨率，但是矩形窗的旁瓣峰值较大，因此其频谱泄露比较严重。

输入语音流采用单声道、8bit、16KHz 采样（实际上语音采样频率为 44.1KHz，单声道，语音保存为.wav 格式，编码比特数 16bit。其中数据库一为较纯净的语音样本，)。 以 256 个采样点为一个音框单位（帧），以 128 为音框之间的重迭单位，对输入语音流进行分帧。计算各帧语音数据的累积能量$\mathrm{E}=\sum_{\mathrm{n-1}}^{\mathrm{N}}\mathrm{x^2(n)},N=256$（最大值为 256^3=16777216，用 int 表示足够），如果连续语音帧累积能量大于预设静音阈值（连续数>100），则采纳该段连续语音帧为训练语音；保留所有可供训练的语音。

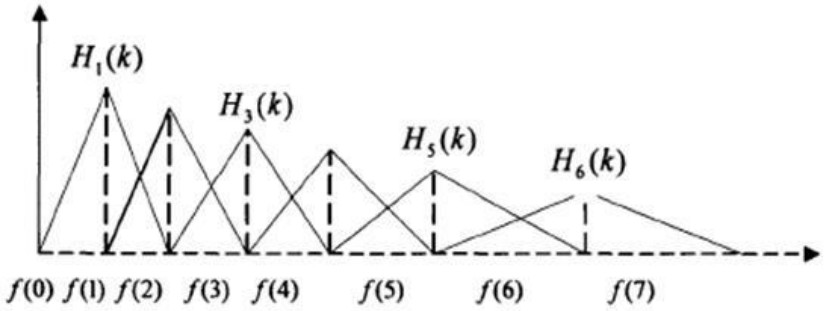
2 MFCC 系数（特征提取）

MFCC 特征系数是一种基于听觉感知频域倒谱系数，研究证明，声音梅尔域（Mel）频率同其物理频率间并不是线性关系，而是在一定范围内表现出对数关系。将语音信号通过一系列三角形的 Mel 带通滤波器组，使得语音更加贴近人耳的非线性感知特性。

2.1 对处理后的语音信号进行离散傅里叶变换（实际设计过程中为了加快运算速度，使用基 2-FFT）

$$\mathrm{X(k)}=\sum_{\mathrm{n=0}}^{\mathrm{N-1}}\mathrm{x(n)\,e^{-j2\pi k/N}},0\le k\le N$$

2.2 把 2.1 中计算获得的频谱系数$\mathrm{X(k)}$经过 Mel 三角带通滤波器组进行滤波处理过程，三角形滤波器组的个数一般都取 20~28 个，本文通过实验选取 40 个。



Mel 频率滤波器组

人耳有一些特殊功能，使其在嘈杂环境及各种变异环境下仍能正常分辨出各种语音，其中耳蜗起很关键的作用，它相当于一个滤波器组，使滤波作用在对数频率尺度上进行，在 1kHZ 以下为线性尺度，1kHZ 以上为对数尺度，使人耳对低频信号比高频信号更敏感。基于这一特点，根据心理学实验得到类似于耳蜗作用的一组滤波器组，即 Mel 频率滤波器组。对频率轴的不均匀划分是 MFCC 区别与倒谱的最重要特点。公式为：$\mathrm{mel(F)} = 2595 * \log_{10}\left(1 + \frac{\mathrm{f}}{700}\right)$。

将语音频率划分为一系列三角形的滤波器，即 Mel 滤波器组，如图。将频率变换到 Mel 域后，Mel 带通滤波器组的中心频率均匀排列。将两个相邻的三角形滤波器的中心频率映射到 Mel 频率上时，其跨度是相等的，也就是对于中心频率$f(i)$，满足：

Mel[f(i+1)]-Mel[f(i)]=Mel[f(i)]-Mel[f(i-1)]

三角形滤波器的两条边表示加权系数，且其频率响应应定义为：

$$H_m(k) = \begin{cases} 0 & ,k < f(m-1) \\ \frac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m) \le k \le f(m+1) \\ 0 & ,k \ge f(m+1) \end{cases}$$

式中$\sum_{m=0}^{M-1} H_m(k) = 1$，$f(m)$表示三角形滤波器的中心频率，$m$表示第$m$个滤波器，$0 \leqslant m \leqslant M$，$M$为总的三角滤波器个数。（注：DFT 是对 DT FT 的频域进行采样，即令 w=2*pi*k/n，相当于在 z 平面 DT FT 的那个单位圆上抽样取点，逆时针分成 n 等份，k 就是第 k 个点，所以很明显 X（k）是周期的）

2.3 对通过三角滤波器组进行滤波后的频谱系数求对数能量

将每帧的频谱参数通过一组 N 个三角形带通滤波器(N 一般为 20~30 个)所组成的梅尔刻度滤波器，将每个频带的输出取对数，求出每一个输出的对数频带能量(log energy)，k = 1,2,… N。

$$\mathrm{s(m)} = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), 0 \le \mathrm{m} \le \mathrm{M}$$ (应该修改这部分代码)

这一步实现了语音信号的同态处理(**同态处理方法是一种设法将非线性问题转化为线性问题来进行处理的方法，它能将两个信号通过乘法合成的信号，或通过卷积合成的信号分开。对于语音信号，我们的目的是要从声道冲激相应与激励分量的卷积中分开各原始分量。由卷积结果求得参与卷积的各个信号分量是涉**

及数字信号处理理论的一项任务，称为“解卷积”或简称“解卷”。对语音信号进行同态分析后，将得到语音信号的倒谱参数，因此同态分析也称为倒谱分析或同态处理），有效的滤掉语音信号中的卷积信道噪声。语音信号可用一个线性时不变系统的输出表示，即看作声门激励信号与声道冲激响应的卷积。

另（优化方案）：在对数能量计算过程中，一帧的音量（即能量），也是语音的重要特征，而且非常容易计算。因此，通常再加上一帧的对数能量（定义：一帧内信号的平方和，再取以 10 为底的对数值，再乘以 10）使得每一帧基本的语音特征就多了一维，包括一个对数能量和剩下的倒频谱参数。（若要加入其它语音特征以测试识别率，也可以在此阶段加入，这些常用的其它语音特征包含音高、过零率以及共振峰等）。

2.4 同态处理后的信号 s(*m*)进行离散余弦变换（DDCT）获得 MFCC(计算获得的序列参数{*c*0,*c*1,……,*cn*}就是某一帧语音帧的 MFCC 系数，其中*c*0 是直流成分，所以舍去不用。)：

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2,\dots,L$$

离散傅里叶变换需要进行复数运算，尽管有 FFT 可以提高运算速度，但在图像编码、特别是在实时处理中非常不便。离散傅里叶变换在实际的图像通信系统中很少使用，但它具有理论的指导意义。根据离散傅里叶变换的性质，实偶函数的傅里叶变换只含实的余弦项，因此构造了一种实数域的变换——离散余弦变换(DCT)。通过研究发现，DCT 除了具有一般的正交变换性质外，其变换阵的基向量很近似于 Toeplitz 矩阵的特征向量，后者体现了人类的语言、图像信号的相关特性。因此，在对语音、图像信号变换的确定的变换矩阵正交变换中，DCT 变换被认为是一种准最佳变换。在近年颁布的一系列视频压缩编码的国际标准建议中，都把 DCT 作为其中的一个基本处理模块。

对上一步所求得的对数能量进行离散余弦变换求出 L 阶的 Mel-scale Cepstrum 参数。L 阶指 MFCC 系数阶数，通常取 12-16。这里 M 是三角滤波器个数。研究表明，最前面若干维及最后若干维的 MFCC 对语音的区分性能较大，因而语音识别中通常只取前 12 维 MFCC（取了 13 维）。

研究表明，在 MFCC 特征中加入 MFCC 差分系数作为声纹识别系统的特征参数能有效的提高识别率。MFCC 一阶差分系数由式：（待优化）

$$d(n) = \frac{\sum_{i=-k}^{k} i \cdot c_{n-1}}{\sqrt{\sum_{\Sigma=-k}^{k} i^2}}$$

计算得到。式中 k为常数，取值为 2，且由此公式还可获得更高阶的差分系数， 可以在声纹密码锁中，提取 MFCC 特征参数时是基于 MFCC 及其一阶差分系数 的，其特征参数共有 24 维，含 12 维的 MFCC 及 12 维的 MFCC 一阶差分参数。

理想 MFCC 参数(待优化)：N 维 MFCC 参数（N/3 MFCC 系数+ N/3 一阶差分参数+ N/3 二阶差分参数）+帧能量（此项可根据需求替换）

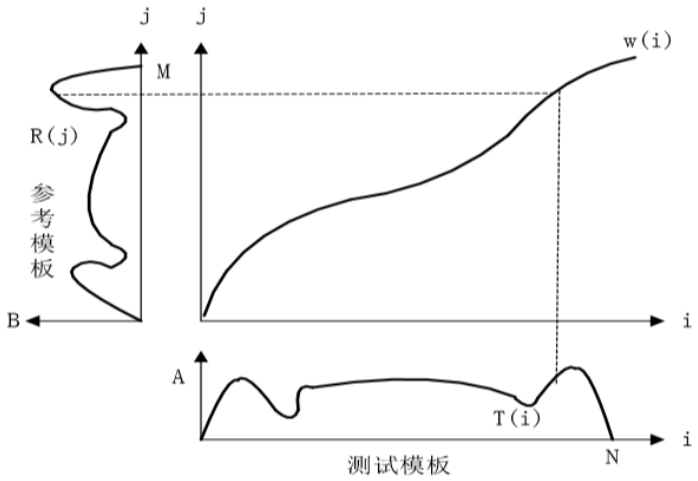
3 模式识别方法（DTW）

模板匹配法，即在规定了说话内容的前提下，在训练阶段，提取训练语音的特征参数序列，作为模板。识别时，提取测试语音的特征参数序列，并与相应的参考模板进行距离计算和模式匹配进行识别。常用的模板匹配方法有以下两种：

（1）动态时间规整（DTW）（选择了这个）；（2）矢量量化方法（VQ）。一般与文本有关的识别采用 VQ，将输入特征序列逐个与 VQ 各码本中的码字比较，然后将距离累加作为识别的依据，而不考虑时序。

3.1 动态时间规整的方法

语音识别中，不能简单地将输入参数及相应的参考模板直接比较，因为语音信号由相当大的随机性。为此，一种简单的方法是对未知语音信号进行伸长或缩短，直至与参考模板的长度相一致；即匹配时对特征向量序列进行线性时间规整。该方法的精度取决于端点的检测精度。因而需要采用非线性时间对准算法。Itakura 将动态规划算法用于解决语音识别中语速多变的难题，提出了著名的动态时间规整算法。DTW 算法就是在时间轴上对两个特征模板进行非线性的弯曲，解决长度不匹配的问题，其核心思想是通过将时间规整和欧氏距离相结合，最后求取两个待匹配矢量的最佳匹配路径和距离。它是效果最好的非线性时间规整模式匹配算法。规整过程中，输入的是两个时间函数，典型的有幅度、共振峰或 LPC 系数。DTW 非线性弯曲原理如图所示：



其中纵坐标表示有 M 帧矢量的参考模板，横坐标代表有 N 帧矢量的测试模板，弯曲的对角线表示二者之间的映射光系。

如设测试语音参数有 N 帧矢量，参考模板有 M 帧矢量，且$\mathrm{N} \neq \mathrm{M}$，则 DTW 就是寻找一个时间规整函数$\mathrm{j} = \mathrm{w(i)}$，将测试矢量的时间轴 i 非线性的映射到模板的时间轴上，并使 w(i)满足：

$$\mathrm{D} = \min_{w(i)}\{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{d}[\mathrm{T(i)}, R(w(i))]\}.$$

式中，$d[T(i), R(w(i))]$ 表示第i帧测试矢量$T(i)$与第j帧参考矢量$R(j)$间的欧氏距离，由式$\mathrm{d[T(i),R(j)]} = \sum_{\mathrm{k}=1}^{\mathrm{l}}(\mathrm{t_k} - \mathrm{r_k})^2$. 中$\mathrm{t_k}$为测试矢量$T$第$i$帧的第$k$个数据，$\mathrm{r_k}$为参考矢量$R$第$j$帧的第$k$个数据，$l$为矢量参数各帧的数据个数，按照所采取的 MFCC 参数个数确定l（采用的是 13）。DTW 不断地计算两矢量间的距离以寻找最优匹配路径，得到的是两矢量匹配时累积距离最小的规整函数，从而保证了二者之间有最大的声学相似特性。由 DTW 算法计算的参考矢量与测试矢量之间的最佳匹配路径如式所示：

$$[\mathrm{T(i)}, \mathrm{R(w(1))}], [\mathrm{T(i)}, \mathrm{R(w(2))}], \ldots, \left[\mathrm{T(i)}, \mathrm{R}\left(\mathrm{w(i)}\right)\right].$$

根据局部性限制，DTW 算法将匹配路径限定在一平行四边形内，找到从 (1,1) 到(N,M)的路径 j = $w(i)$使路径上的值累计距离最小，如图所示：

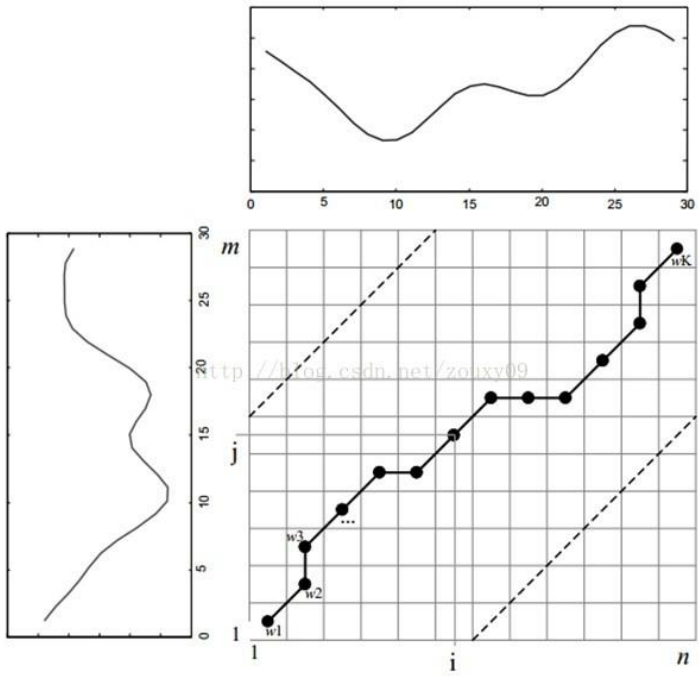


Figure 3: An example warping path.

首先，路径不是随意选择的，需要满足以下几个约束：

1）边界条件：D1=(1, 1)和 DK=(m, n)。所选的路径必定是从左下角出发，在右上角结束。

2）连续性：DTW 不可能跨过某个点去匹配，只能和自己相邻的点对齐。

3）单调性：D 上面的点必须是随着时间单调进行的。以保证中的虚线不会相交。

结合连续性和单调性约束，每一个格点的路径就只有三个方向了。例如如果路径已经通过了格点(i, j)，那么下一个通过的格点只可能是下列三种情况之一：(i+1,j)，(i,j+1)或者(i+1, j+1)。

总代价函数为

$$D[c(k)] = d[c(k)] + \min D[c(k-1)] = d[c(k)] + \min \begin{pmatrix} D & (n-1,m) \\ D & (n-1,m-1) \\ D & (n,m-1) \end{pmatrix}$$

(注：可优化 DTW 算法，修改 Test 文件加权评分公式)

参考文献：

[1] 周颖 武汉理工大学 Android 声纹密码锁设计 2014.5

[2] 语音特征参数 MFCC 提取过程详解

https://my.oschina.net/jamesju/blog/193343

[3] 文本无关的声纹识别 验证

https://blog.csdn.net/c395565746c/article/details/6210920

[4]胡航 现代语音信号处理 电子工业出版社 2014.7