实

验

二

三

实

验

报

告

小组成员：王少崎、曾泽源、魏育豪、郑纯然、郭广帅

语音信号的频域特征分析和基于动态时间规整(DTW)的孤立字语音识别实验

**摘要**

关键词：

1. **引论**

随着现代科学和计算机技术的发展，人们在与机器的信息交流中，需要一种更加方便、自然的方式。而人的直观感觉可以给人最直接的印象，获取信息速度也就最快。虽然，嗅觉、触觉也是人类固有的感觉，人们可以从中得到某些外界信息，但最重要、最精细的信息源只有图像和语言两种。而且，语言是人类最重要的、最有效的、最常用的和最方便的通信形式。这就很容易让人想到能否用自然语言代替传统的人机交流方式如键盘、鼠标等。

相较于时域语音识别，频域语音识别更能反映语音的物理特性，从而更精确地完成语音识别任务。而70年代DTW基本成熟，解决了Mel频率谱计算得到的频谱特征向量长度各不相同的问题。

1. **实验原理**

**2.1.**短时傅里叶变换

设语音信号时域信号为,加窗分帧处理后得到的第帧语音信号为 ,则满足下式

设离散时域采样信号为,其中为时域采样点序号, 是信号长度,然后对信号进行分帧处理,则表示为,其中是帧序号,是帧同步的时间序号。信号的短时傅里叶变换为

定义角频率,则得离散的短时傅里叶变换(DFT),它实际上是在频域的取样,如下所示:

在语音信号数字处理中,都是采用的离散得里叶变换来替代,并且可以用高效的快速傅里叶变换(FFT)算法完成由至的转换。当然,这是窗长N必须是2的倍数(L是整数),以便利于实现按间抽取或按频率抽取的蝶形运算。

**2.2.**Mel频率倒谱系数

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：

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

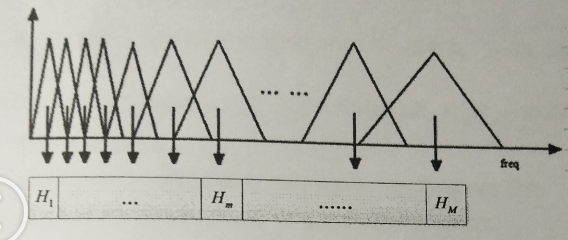


图2.1Mel频率尺度滤波器

设划分的带通滤波器为,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,在Mel频率范围内。这些滤波器是等带宽的(如图2.2)

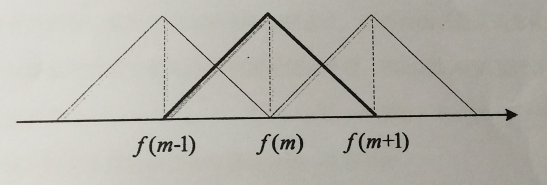


图2.2相邻三角滤波器之间的关系

每个带通滤波器的传递函数为

其中,

Mel滤波器的中心频率定义为

其中,和分别是滤波器组的最高频率和最低频率,为采样频率,单位为,是滤波器组的数目,为FFT变换的点数,。

**2.3.**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。

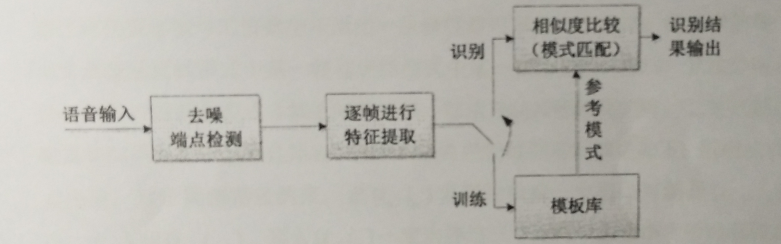


图3.1模板匹配法语音系统的原理框图

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。

**2.4.**动态时间规整

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,要求待比对的特征间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,假设词库中输入语音的特征矢量序列为输入语音的特征矢量序列为,那么动态时间规整就是要找到时间规整函数。该函数把输入模板的时间轴非线性映射到参考模板的时间轴m，并满足下式：

式中,表示两帧矢量之间的距离；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：

1. **相关工作**

**3.1.**根据短时傅里叶变换的原理,编写其函数,自行实现FFT过程

**3.2.**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2.3MFCC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,式中,是一个常数,如0.97。

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。

[3]、计算谱线能量

计算每一顿FFT后的数据谱线能量：

[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(其中表示第帧,表示频域中的第k条谱线)与Mel滤波器的频域响应相乘并相加：

[5]、经离散余弦变换(DCT)得到MFCC系数

式中,是(4)求出的Mel滤波器能量；是指第个Mel滤波器（共有个）；是指第帧；是MFCC系数阶数,通常取12-16。

**3.3.**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顿之间的距离,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,点可到达的前一个格点可能是，和,那么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：

这样从(1,1)点出发搜索,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,得到最小匹配距离所对应语音即为识别结果。

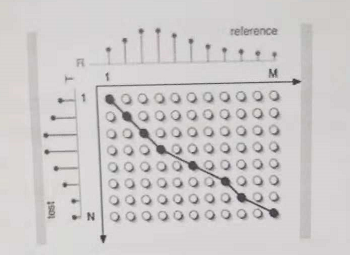
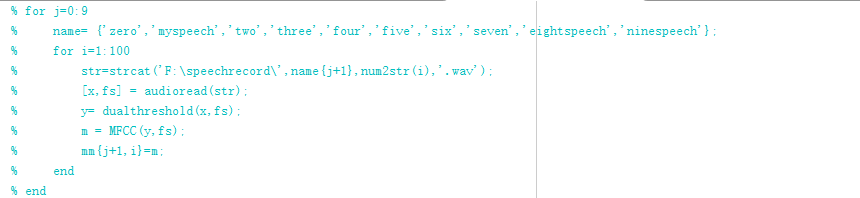


图3.2 DTW算法原理图

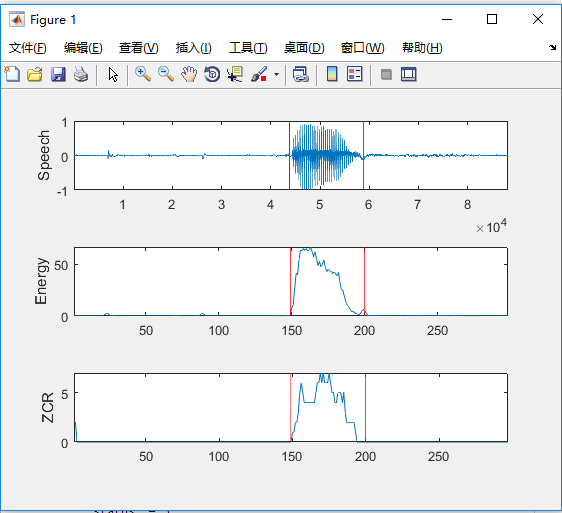
**3.4.**面向“0”-“9”这10个孤立语音的识别任务，实现频域法语音识别，各个功能模块均采用编程来实现，包含必要的界面，能够自动地完成语音识别的完整过程。

1. **实验过程**

**1.提取语音信号，通过双门限法进行预处理，并提取MFCC特征。（并将计算出的mm save下来方便后续的计算）**

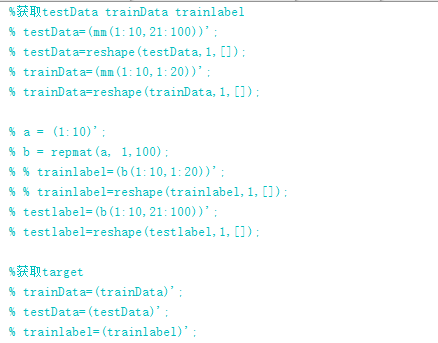


**双门限运行效果:(对于单个语音信号截取效果如下，左右红线内即为所得)**

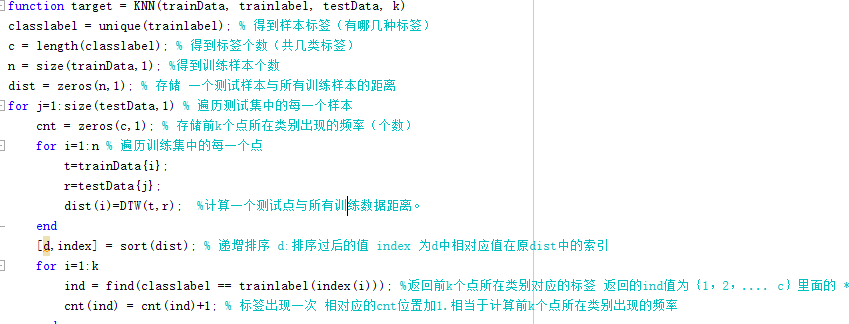


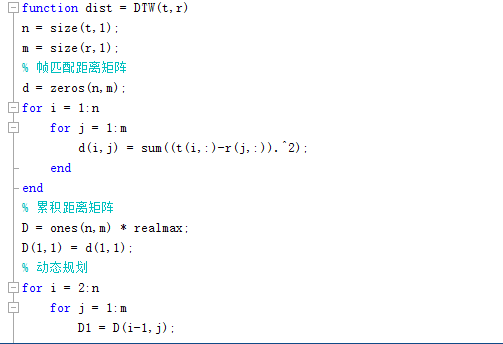
**2.获取testData,testlabel,trainData**

**以每个语音信号前20个构成训练集，剩下的总共800语音信号作为测试集。(也是依次save，节省运算时间)**



**3.调用KNN 函数，以DTW计算出的大小为准，以前15小的距离中标签数最多的最为测试集的标签**

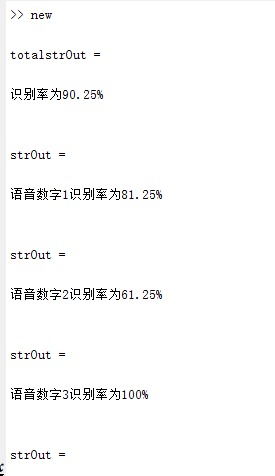


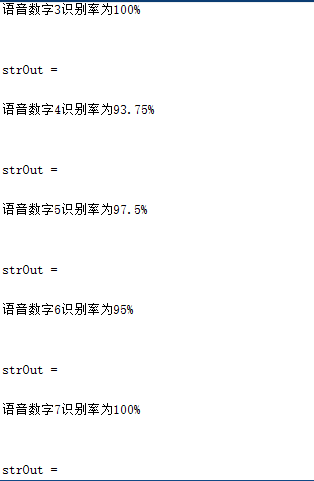
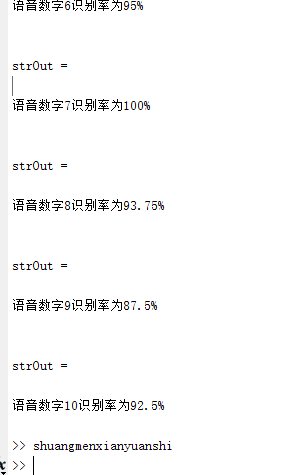


1. **计算总体的识别率和单个数字的识别率**



1. **实验结果及分析**



整体的识别率达到了90.25%，说明利用频域特征进行语音识别的效果还是要远远优于时域粗略的特征。