基于频域分析的语音识别*

赵 超, 庞 立, 彭宣尧, 龙雅琴, 林逸阳
（西安交通大学 自动化科学与工程学院, 陕西 西安 710049）

摘 要 本文对用户输入的孤立数字和名字语音频域的频域特性展开分析，以达到基于语音频域特性的语音识别目的。在基于matlab为实验工具的条件下，先将预处理后语音进行提取MFCC频域特征参数，最后基于动态时间规整技术对孤立数字和名字进行模板批匹配。最后可以得到频域的识别准确率明显高于基于时域参数识别的准确率，而且在不同窗函数下的预处理语音信号也会显著影响语音识别的准确率。

关键词 域特征，语音识别，DTW

Speech Recognition Based on Frequency Domain Analysis

Chao Zhao,Li Pang,Xuanyao Peng,Yaqin Long,Yiyang Lin
(College of Automation Science and Technology, Xi'an Jiaotong University, Xi'an Shaanxi 710049, China)

Abstract In this paper, the frequency domain characteristics of isolated numbers and names input by users are analyzed in order to achieve the purpose of speech recognition based on the frequency domain characteristics of speech. Under the condition of using MATLAB as the experimental tool, the preprocessed speech is first extracted from the MFCC frequency domain characteristic parameters, and finally the template batch matching of isolated numbers and names is carried out based on the dynamic time warping technology. Finally, the recognition accuracy in frequency domain is significantly higher than that based on time domain parameters, and the preprocessed speech signals under different window functions will also significantly affect the accuracy of speech recognition.

Key Words domain features, speech recognition, dtw

0 引言

在做了基于语音时域特征进行语音识别之后，由于时域特征的局限性-时域分析对语音信号的频率特性没有直观的了解，为了提高准确率，我们进行基于频域分析的语音识别，希望在新的方向提高对孤立数字和名字的识别。

*收稿日期: XXXX−XX−XX. 基金项目：国家自然科学基金资助项目 (51685168)

1 实验原理

1.1 FFT

1.1.1 FFT基本原理

有限长序列x（n）的N点DFT为如下定义：

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{kn} \tag{1}$$

显然，直接DFT运算概念清楚、编程简单，但占用内存大、运算速度低，在实际工作中性价比极低。而基2-FFT算法的基本思想是把原始的N点序列依次分解成一系列短序列，充分利用旋转因子的周期性和对称性，分别求出这些短序列对应的DFT，再进行适当的组合，得到原N点序列的DFT，最终达到减少运算次数，提高运算速度的目的。

按时间抽取的基2-FFT算法，先是将N点输入序列$x(n)$在时域按奇偶次序分解成$\frac{N}{2}$个点序列和$x_1(n)$和$x_2(n)$，再分别进行DFT运算，求出与之对应的$X_1(k)$和$X_2(k)$，然后利用1所示的运算流程进行蝶形运算，得到原N点序列的DFT。

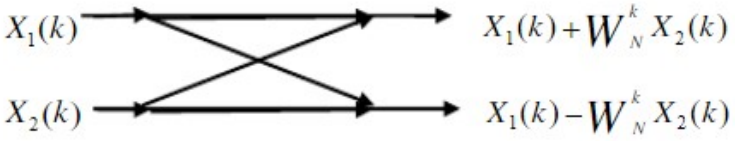

图 1: 蝶形运算

只要N是2的整数次幂，这种分解就可一直进行下去，直到其DFT就是本身的1点时域序列。一个完整的8点DIT-FFT运算流程如图**??**所示。

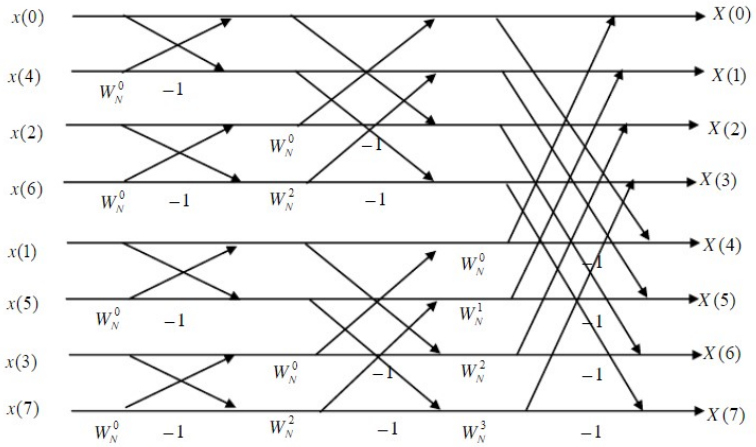

图 2: 8点DIT-FFT运算流程

1.1.2 DIT-FFT编程框图

MATLAB提供的FFT函数是一个计算DFT的智能程序，能自动选择快速算法进行DFT运算，由于它是一个内建函数，用type命令无法直接查看其程序代码。

MATLAB的数组元素按序存储，可用下标寻访，但下标是从1开始的，所以在MATLAB程序中，寻访数组中的元素(数据)时，下标要在原序号上加1。旋转因子可按指数预先计算出来并存放到数组W_n中，虽然占用了一些内存，但程序运行时可直接寻访调用，无需反复计算，可进一步提高运算速度。因此，用MATLAB实现DIT-FFT算法的程序框图如图3所示[1]。

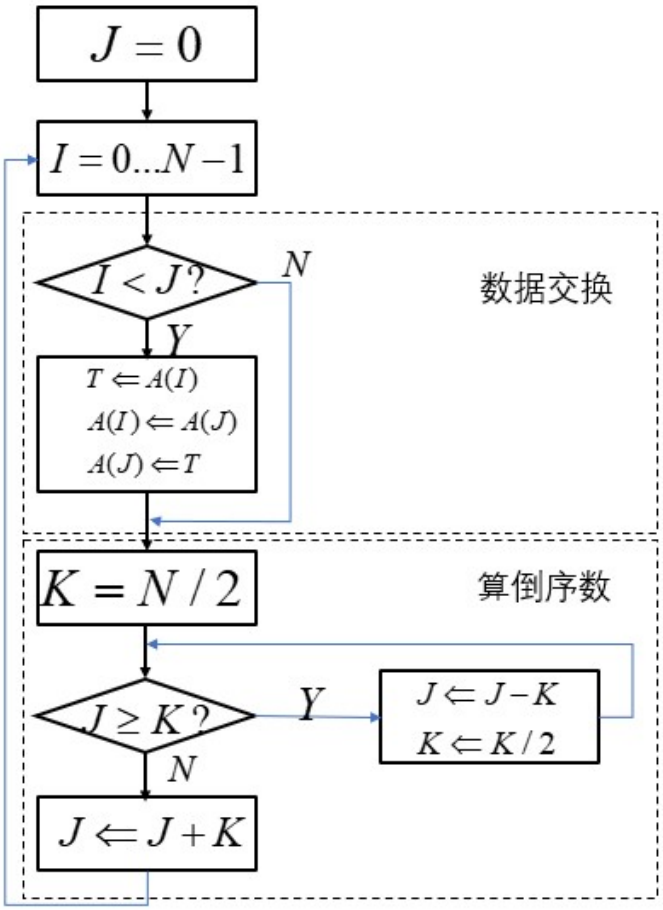

图 3: matlab实现FFT程序框图

1.2　Mel谱和MCFF

人的听觉系统是一个较为特殊的非线性系统，它相应不同频率信号的灵敏度是不同的。在语音信号特征的识别上，人耳的听觉系统不仅可以提取出语音中的语义信息，也可以提取出说话者的个人特征。因此，模拟人耳的听觉感知处理方式，可以大大提高语音的识别率。

梅尔（Mel）频率分析便是基于人类听觉感知实验的。实验中观测到，人耳其实就像一个滤波器组，只关注某些特定的频率分量，即人的听觉对频率是有选择性的。而人耳中的滤波器在频率坐标轴上不会统一分布，在低频区有很多滤波器，分布较为密集，而在高频区，滤波器的数目大大减少，且分布很稀疏。因此，梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MCFF)考虑到了人耳的听觉特征，先将线性频谱映射到基于听觉感知的Mel非线性频谱中，再转换到倒谱中。

通过实验得知，Mel频率尺度的值大体对应实际频率的对数分布关系，与实际频率的关系可用下式近似表示：

$$Mel(f) = 25951lg(1 + f/700) \tag{2}$$

其中，f为频率，单位为Hz。其关系图可用图4表示：

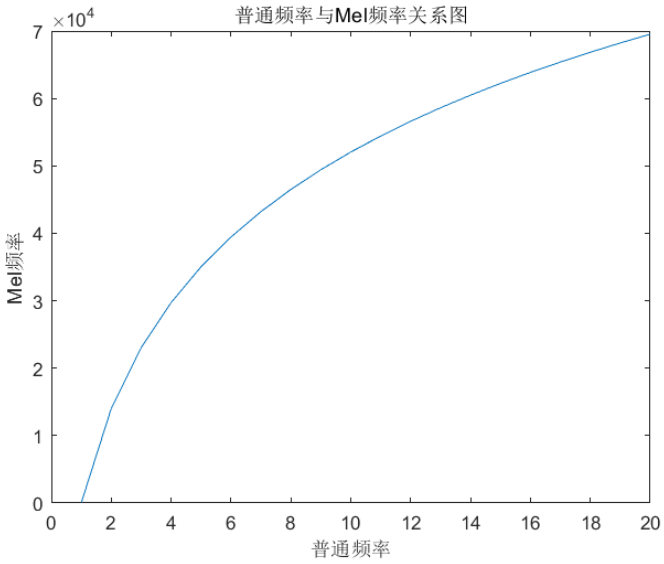

图 4: 普通频率与Mel频率关系图

语音频率可以被划分成一系列三角形的滤波器序列，也就是Mel滤波器组。设划分的带通滤波器为$H_m(k)$,$(0 \le m \le M$,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，这些滤波器都是等宽的。

每个带通滤波器的传递函数为：

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le\le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{3}$$

其中，$\sum_m^{M-1} H_m(k) = 1$。

Mel滤波器的中心频率f(m)定义为：

$$f(m) = \frac{N}{f_z} F_{Mel}^{-1}(F_{Mel} f_i + m \frac{F_{Mel}(f_h) - F_{Mel}(f_i)}{M+1}) \tag{4}$$

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

1.3 DTW

Dynamic Time Warping(DTW)技术是基于动态规划思想产生的，它可以实现不等长特征向量的距离计算。动态时间规划用满足一定条件的时间规整函数来描述输入模板和参考模板的对应时间关系，求解两模板匹配时累计距离最小所对应的规整函数。假设词库中某一参考模板的特征向量列为$a_1, ..., a_m, ..., a_M$，输入语音的特征向量列为$b_1, ..., b_n, ..., b_N, M \ne N$，那么动态时间规整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，并满足下式：

$$D = min \sum_{n=1}^{N} d[n, T(n)] \tag{5}$$

式中，d[n,T(n)]表示两帧向量间的距离，D是最佳时间路径下两个模板的距离测量。在本实验中，本文采取欧氏距离：

$$d(x,y)=\frac{1}{k}\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \tag{6}$$

2　实验数据

我们通过matlab中*audiorecorder*()函数录制得到原始音频数据，音频的长度为两秒钟，采样频率为44100Hz，采样位数为16位，采样通道数为2个通道，每个人每个数字录制两次，总共得到0到9的数字原始数据的数量各20个；同时得到五个成员的名字语音数据各4个。

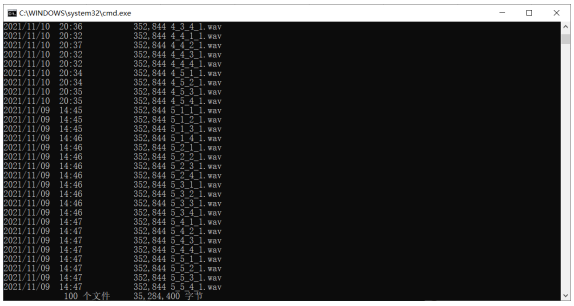

图 5: 名字音频信息

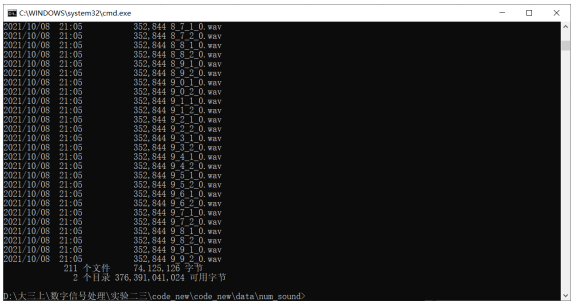

图 6: 数字音频信息

3　实验部分

实验部分的流程为：

1) 通过电脑录制音频；

2) 将得到的音频数据进行预处理，经过端点检测得到理想的音频数据；

3) 提取音频数据的MFCC特征；

4) 进行DTW算法搜索；

5) 将得到的结果进行总结归纳；

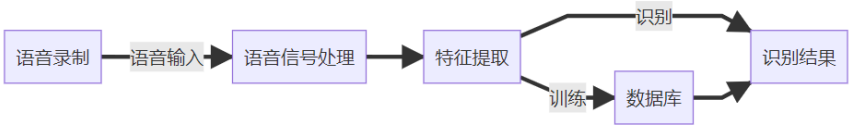

图 7: 实验流程

初始短时能量高门限	50
初始短时能量低门限	10
初始短时过零率高门限	10
初始短时过零率低门限	2
最大静音长度	8ms
语音最小长度	20ms

表 1: 端点检测的参数指标

3.1 端点检测

由于我们录制的环境不是绝对安静的，每个人的录音设备也不尽相同，所以每个人录制的音频，所以会出现带有一定噪音的音频，为了从带有噪声的语音中准确的定位出语音的开始点，和结束点，去掉静音的部分，去掉噪声的部分，找到一段语音真正有效的内容，我们进行端点检测（Voice Activity Detection），这里我们进行的是基于阈值的VAD，通过提取时域（短时能量、短期过零率等）或频域（MFCC、谱熵等）特征，通过合理的设置门限，达到区分语音和非语音的目的。

我们的端点检测初始特征阈值如表1所示，通过这样的指标，我们检测出来的音频信息可以过滤掉大部分的噪音，双端检测的结果如图18和图24所示。

我们将端点检测得到的音频信息保存为txt文件格式方便后续使用（图25），同时，为了数据的准确性，我们对音频进行加窗后再次进行端点检测，分别得到音频在加入矩形窗、汉宁窗、汉明窗后双端检测得到的数据，使得数据的处理更加完善。

3.2 MFCC特征提取

3.2.1 分帧

我们将所得到的数据进行分帧，语音整体上是非平稳信号，我们取256个数据点为一帧，为避免相邻帧间的过大变化，帧之间应设置重叠，选取帧移为帧长的1/3，即80帧。另外为了方便后面的FFT计算，这里的帧长取为2的整幂指数。

3.2.2 傅里叶变换

语音的时域信号变化迅速，不易分析，而频域变化相对缓慢，因此对加窗和端点检测完成后的信号进行快速傅里叶变换FFT，再折半舍弃对称部分后平方计算信号能量谱。这里我们对每一帧语音进行256点的快速傅里叶变化。图36列举了数字0-9第一帧FFT的图形。

3.2.3 三角带通滤波器

将每帧能量谱中的线性频谱刻度转化成符合人耳听觉特性的梅尔刻度后，将其通过如图37的一组24个三角形带通滤波器，计算通过每个频带的能量并取对数。

3.2.4 离散余弦转换

离散余弦转换(DCT)系数是用来将能量集中在前面几项中，达到在减少判别参数提高运算速度的同时又不失其准确性的目的。公式7即为对数能量E_k转换成L阶的梅尔频率倒谱系数。

$$\sum_{k=1}^{N} E_k cos(m(k-\frac{1}{2}\frac{\pi}{N})), m = 1, 2, 3 \cdots L \tag{7}$$

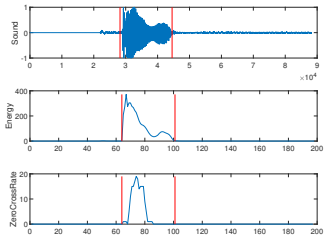

图 8: 数字0端点检测结果

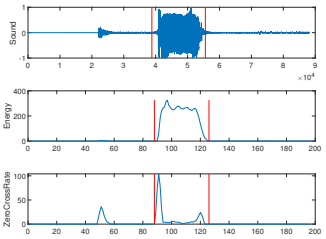

图 9: 数字1端点检测结果

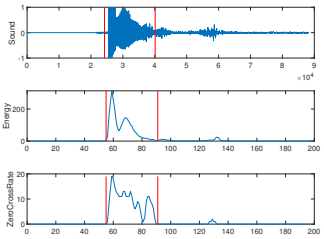

图 10: 数字2端点检测结果

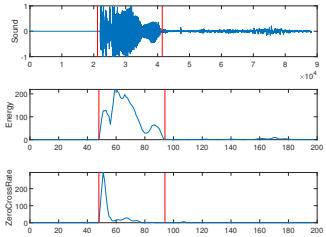

图 11: 数字3端点检测结果

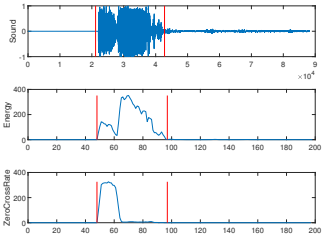

图 12: 数字4端点检测结果

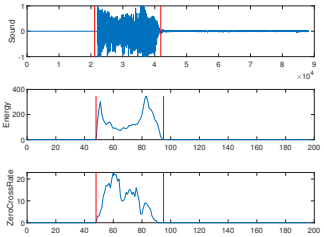

图 13: 数字5端点检测结果

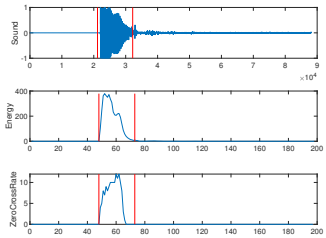

图 14: 数字6端点检测结果

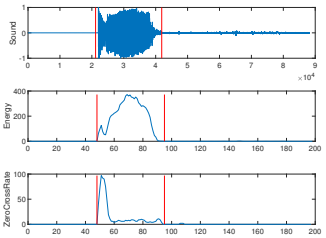

图 15: 数字7端点检测结果

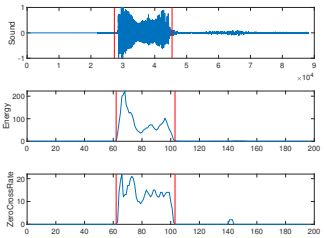

图 16: 数字8端点检测结果

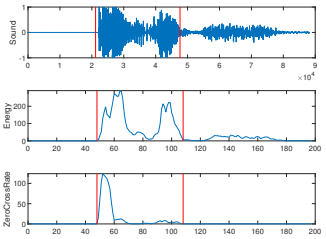

图 17: 数字9端点检测结果

图 18: 数字端点检测结果

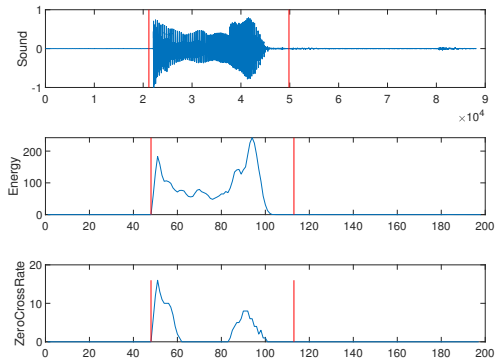

图 19: 名字“庞立”端点检测结果

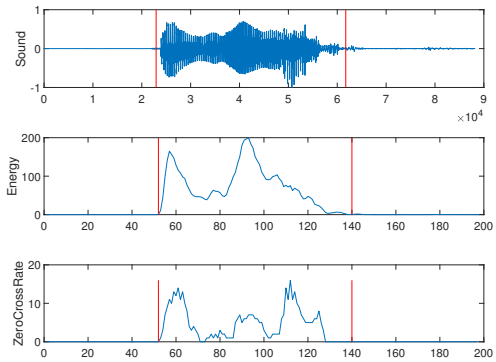

图 20: 名字“林逸阳”端点检测结果

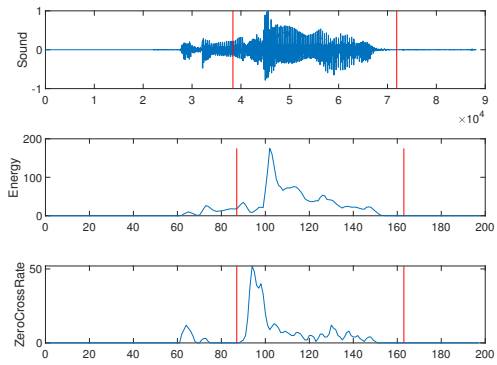

图 21: 名字“彭宣尧”端点检测结果

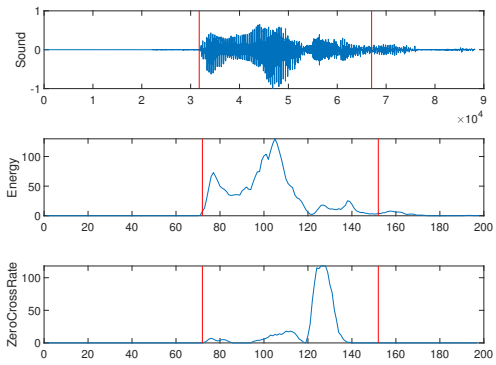

图 22: 名字“龙雅琴”端点检测结果

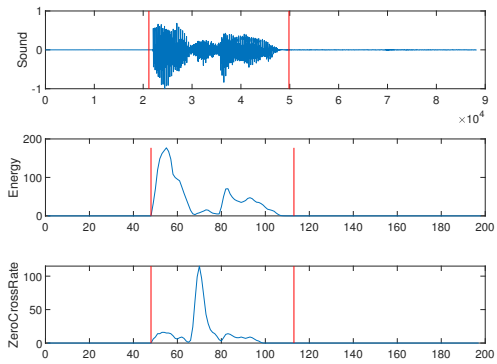

图 23: 名字“赵超”端点检测结果

图 24: 名字端点检测结果

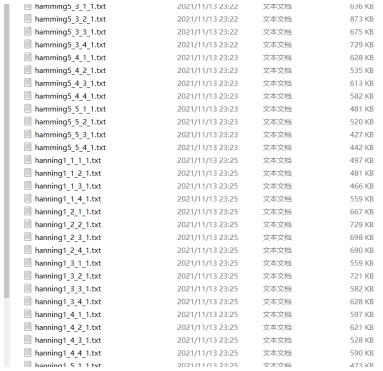

图 25: 端点检测数据

矩形窗下名字准确率	0.9598
矩形窗下数字准确率	0.7237
海宁窗下名字准确率	0.9598
海宁窗下数字准确率	0.7471
汉明窗下名字准确率	0.9598
汉明窗下数字准确率	0.7429

表 2: 准确率

3.2.5 差量倒谱参数

很多文献中提到利用差量倒谱参数来对语音的动态特性参数进行描述[2]，可以提高系统的识别能力。一阶差分MFCC参数表现当前语音帧与前一帧之间的关系，体现帧与帧（相邻两帧）之间的联系，所以我们合并mfcc参数和一阶差分mfcc参数形成MFCC参数，并且去除首尾两帧，因为这两帧的一阶差分参数为0，增加MFCC参数的准确性，增强实验的预测能力。我们将名字“庞立”的一组音频的mfcc参数和一阶差分MFCC参数进行逐帧绘图，结果如38和39所示,

3.2.6 DTW算法搜索

在这里我们使用matlab自带的DTW算法，部分代码如40所示。

3.3 语音识别结果

我们使用matlab的DTW算法对数字和名字的MFCC参数进行搜索，将得到的距离指标保存下来，用交叉验证的方法，训练集与测试集的比例为9:1，在进行1000次的重复试验后，得到的数字识别在不同的窗函数下得到的准确率为不尽相同，如图47所示，x轴为测试次数，y轴为准确率，在矩形窗下数字识别整体准确率最低，而在汉明窗和海宁窗下的数字识别准确率差不多；但是名字准确率在三个窗下的准确率波动不大。具体的平均准确率数据如表2。

4 实验结论

将语音经过预处理，之后通过频域特征分析，即加窗分帧后得到的n帧语音信号进行短时傅里叶变换饿到n帧语音的频域表示形式；之后我们使用Mel频率倒谱系数原理提取语音特征参数MFCC，最后通过DTW对每个孤立语音信号进行模块匹配。在进行多次的实验后发现，DTW技术对孤立语音的识别准确率能够达到70%左右，大大高于时域分析中基于集成学习分类的语音识别效果；其中，对

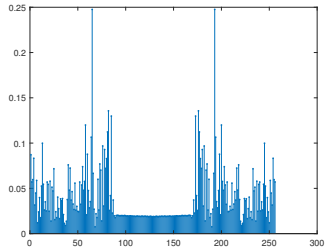

图 26: 数字0FFT结果

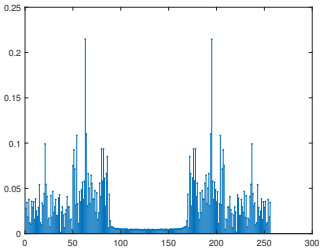

图 27: 数字1FFT结果

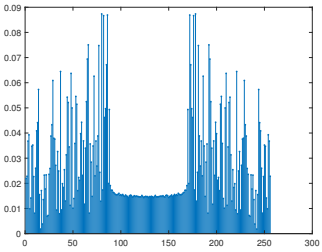

图 28: 数字2FFT结果

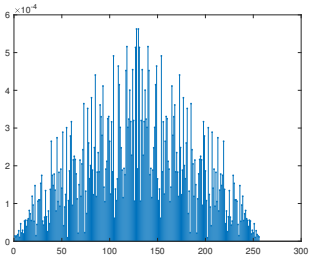

图 29: 数字3FFT结果

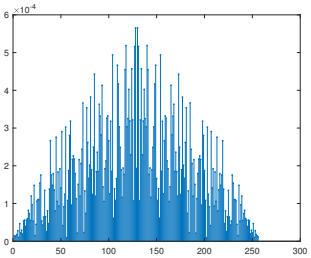

图 30: 数字4FFT结果

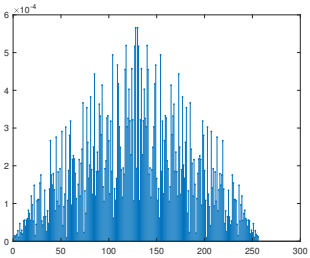

图 31: 数字5FFT结果

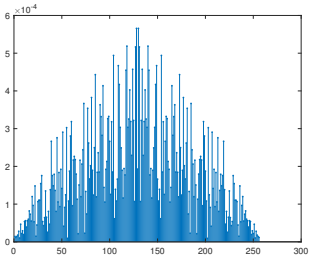

图 32: 数字6FFT结果

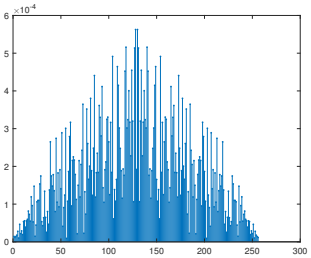

图 33: 数字7FFT结果

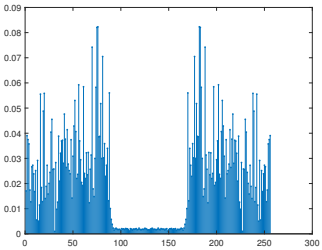

图 34: 数字8FFT结果

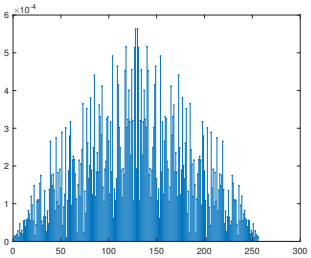

图 35: 数字9FFT结果

图 36: 数字FFT结果

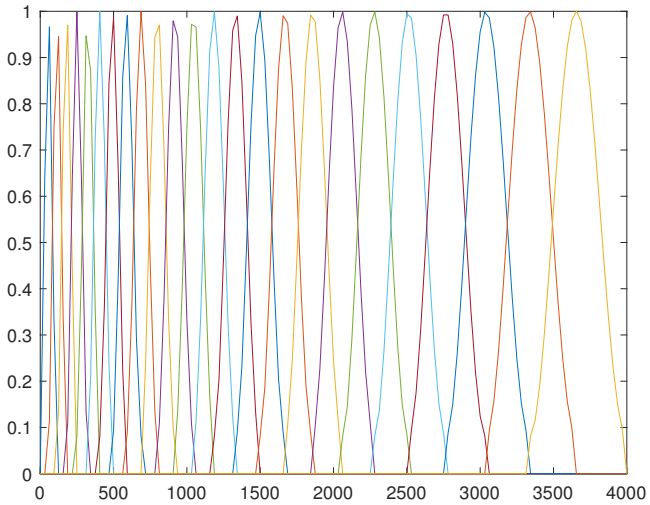

图 37: 梅尔频率滤波器组

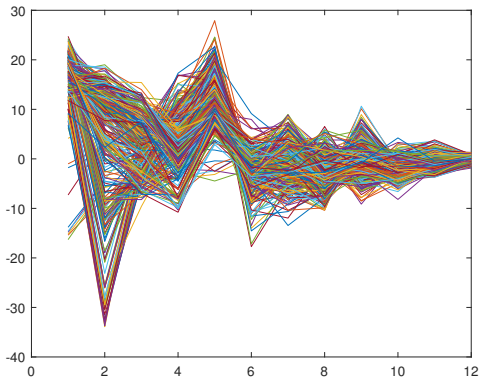

图 38: mfcc参数逐点绘图

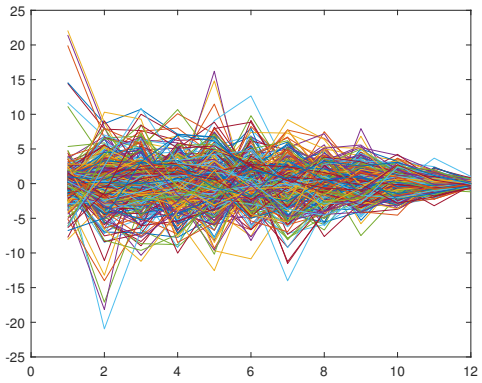

图 39: 一阶差分MFCC参数逐点绘图

```
dis_mat = zeros(length(feature_all), length(feature_all));
for i = 1:length(feature_all)
    for j = 1:length(feature_all)
        dis_mat(i,j) = dtw(feature_all{i}', feature_all{j}');
    end
end

save distance_name_all_mat.txt -ascii dis_mat
save label_name_all_mat.txt -ascii label_all
```

图 40: dtw部分代码

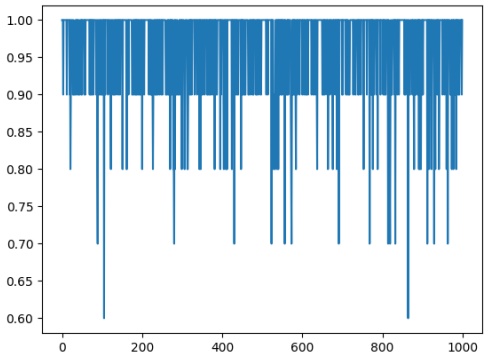

图 41: rectangle name accuracy

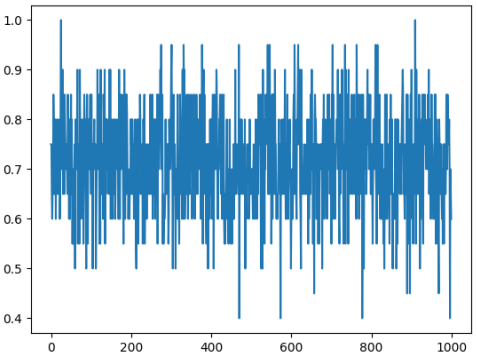

图 42: rectangle num accuracy

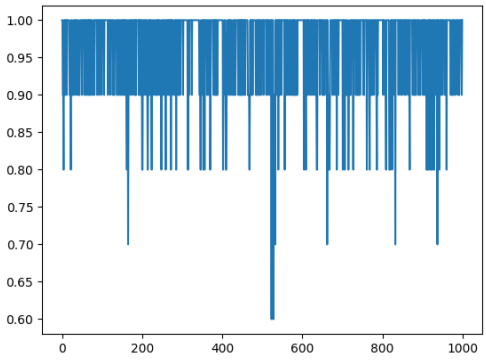

图 43: hanning name accuracy

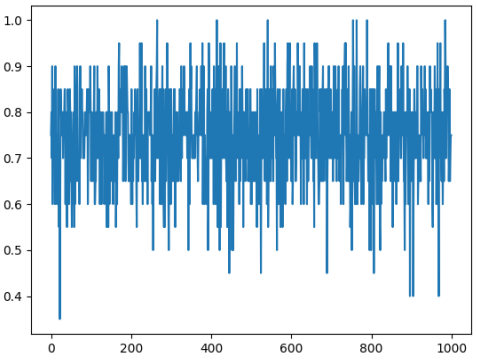

图 44: hanning num accuracy

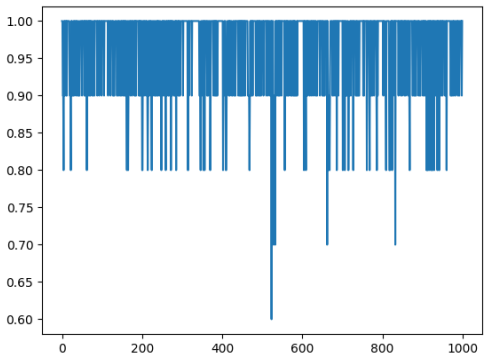

图 45: hamming name accuracy

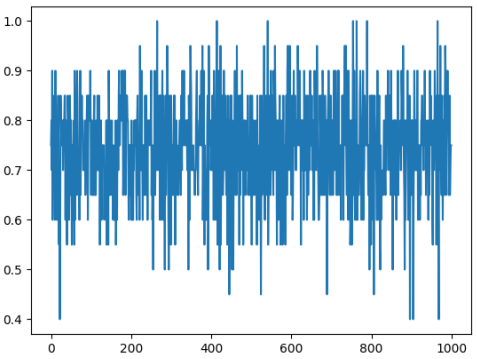

图 46: hamming num accuracy

图 47: 识别结果

语音信号加汉宁窗能够有效提高识别的准确率；另一方面，对于名字的识别基本上可以高达95%，基本上能够识别名字语音信号。

但是我们仍然需要注意的是本次实验存在着语音数据样本不够多的缺陷，导致并不能很好的体现出实验效果。

参考文献

[1] 张登奇,李宏民,李丹.按时间抽取的基2FFT算法分析及MATLAB实现[J].电子技术,2011,38(02):75-77.

[2] 王 昕,张 洪 冉.基 于DNN处 理 的 鲁 棒 性I-Vector说 话 人 识 别 算 法[J].计 算 机 工 程 与 应 用,2018,54(22):167-172.