基于动态时间规整的数字信号语音识别研究

任翌玮，李相宜，马茂原

摘要

本文研究了基于梅尔频谱(MFCC)特征和动态时间规整(DTW)算法的数字语音识别方法。首先介绍了 MFCC 特征提取的过程,包括梅尔刻度、滤波器组和倒谱等技术，以及 DTW 算法的基本原理,它通过在测试模板和参考模板之间寻找一条最优时间扭曲路径来实现时间序列的最佳匹配。然后,建立了一个 0-9 数字语音识别系统。实验采集了 10 个数字的语音样本,提取出 MFCC 特征,并将测试语音与参考模板库中样本逐一匹配,选择匹配距离最小的样本对应的数字作为识别结果。实验结果表明,我们的方法识别性能好，语音识别实时性强。不仅如此，我们还对模型进行了详细的对比实验与消融实验，并且对实验结果进行了丰富的可视化展示。我们还构建了 0-9 的数字识别系统，并且实现了用户友好的语音识别 GUI 界面,用户可以方便快捷地进行实时语音识别项目。

关键词:语音识别;动态时间规整；梅尔频谱

引言

语音识别技术具有重要的理论价值和广阔的应用前景。它融合了数字信号处理、模式识别等多学科知识,可以实现人机交互、语音控制等智能功能。DTW 算法通过在时间轴上扭曲序列信号,寻找测试序

列与参考序列的最优匹配路径,能够抑制序列时间变异的影响,适合处理语音等时间序列信号。而梅尔频谱系数结合了人耳对声音感知特性的梅尔刻度,提取了频谱包络信息。本文将两种技术应用到数字语音识别中,通过模板匹配的方式,验证语音识别算法的效果。不仅如此,本实验还对模型中的各种参数进行了对比实验和消融实验,对实验数据进行了丰富的可视化展示。我们对模型进行了封装打包,设计了用户友好的 GUI 界面,可以方便快捷地进行 0-9 数字的实时语音识别。

算法介绍

一、梅尔频谱系数

1. 梅尔刻度

人耳对声音的感知符合对数关系,即对低频更加敏感。为了建模人耳的非线性特性,在语音分析中常用梅尔刻度[1]取代线性的赫兹频率。梅尔刻度通过一个非线性函数将赫兹频率映射到对应感知频率,其公式为;M(f)=2595log(1+f/700)。其中,f 为线性频率,M(f)为对应的梅尔频率。可以看出,在低频部分,梅尔刻度与赫兹刻度变化趋于线性,而在高频部分,梅尔刻度对频率变化的敏感度下降。梅尔刻度更符合人耳的感知。

2. 梅尔滤波器组

根据梅尔刻度,可以在 300-3700Hz 的语音频带内,按一定间隔制作出一组三角形滤波器,即所谓的梅尔滤波器组。这组滤波器根据人

耳工程模型设计,其带宽也对应于耳朵对声音在不同频率下的分辨率。梅尔滤波器组输出也称为滤波器组谱。一般取 20-40 个滤波器,相邻两个的中心频率之比为一个常数。该滤波器组能够模拟耳朵的频谱分析功能。

3. 倒谱系数提取

信号经过梅尔滤波器组后,通过离散余弦变换,提取各个梅尔频带的短时能量。然后取每个梅尔频带的对数能量值,即为梅尔频谱系数(Mel 频谱系数)。最后取倒谱,即可得到 MFCC 特征。倒谱在抑制信号全局平移的同时,保留了包络信息。因此 MFCC 特征能够保持辨识度的同时,也抑制频谱在轴向的变化,具有一定的鲁棒性。

综上,MFCC 联合梅尔刻度与倒谱技术,提取了信号谱包络信息,模拟了人耳听觉系统的分析过程,成为语音识别领域最常用也最有效的特征之一。

二、DTW 算法原理

DTW 算法[2]的处理流程主要包括四个步骤:

1. 测试模板与参考模板的表示

在一个二维坐标系中,将测试模板的帧序列在横坐标上表示,参考模板的帧序列在纵坐标上表示。如果测试模板有 M 帧{f1,f2,...,fM},参考模板有 N 帧{g1,g2,...,gN},则可以在 M*N 的网格中,用矩阵的行索引和列索引分别对应测试模板和参考模板在时间轴上的帧索引。

2. 帧距离计算

计算测试模板与参考模板在各帧之间的距离,构成一个 M*N 的距离矩阵。常用的距离度量包括欧几里得距离、马氏距离等。依次计算出所有帧对之间的距离,得到完整的距离矩阵。

3. 最优路径搜索

在帧距离矩阵上搜索一条最优时间扭曲路径。该路径从起点(1,1)到终点(M,N),且对应的累积距离最小。在搜索过程中,每前进一步,选择当前点的下、右、右下三个可选点中的距离最小者作为前续点。

4. 累积距离计算

沿最优路径累积所有帧距离和,作为测试模板与参考模板之间的匹配距离。匹配距离最小的参考模板,对应的分类作为识别结果。

上述过程实现了在时间轴上非线性地扭曲测试模板与参考模板序列的匹配,抑制了时间轴变异的影响,达到最优匹配。搜索最优路径的过程可以看作是在帧距离矩阵上定义了一个递推过程,通过动态规划不断逼近最优解。

基于 DTW 和 mel 的数字语音识别方法

基于 DTW 算法的数字语音识别系统主要包括语音信号采集、特征提取、模板训练和 DTW 匹配识别四个模块。

1. 语音信号采集

我们的语音信号采集工作已在实验一的报告中详细介绍。

2. 特征提取

对采集的语音信号帧化,每帧长度为 20ms,帧移 10ms。常用的语音特征包括梅尔频谱系数(MFCC)、线性预测系数(LPC)、声谱分析等。本实验提取的是 MFCC 特征,通过傅里叶变换捕捉谱信息,再映射到梅尔滤波器组上,取倒谱作为特征。该特征抑制了语音信号在频率轴的变化,保留了重要的包络结构信息。

3. 模板训练

从收集的500个语音样本中,选取每个数字词汇中40个表示性的样本,总计 400 个,作为训练样本。针对每个数字利用训练样本计算该类别的平均特征模板。测试时将匹配距离最小的模板所对应的数字作为识别结果。

4. DTW 匹配识别

针对测试语音,与每个数字模板逐一进行 DTW 匹配,计算匹配距离。选择匹配距离最小的模板对应数字作为识别结果。DTW 匹配过程中,测试语音和某数字模板的 MFCC 特征序列,在二维坐标系中分别标记;计算二者之间的欧式距离矩阵;在矩阵上搜索最优路径,使匹配距离最小;最终距离最小的模板对应的数字作为识别结果。通过上述模块的集成,形成了基于 DTW 的数字信号语音识别系统。

消融实验与对比实验

1.mfcc 的维度

首先，我们将不同的 mfcc 特征向量维度进行对比。我们统一选用矩形窗，mel 滤波器数量均为 20 的情况，分别使用 mfcc 特征向量维度为 13, 20 以及 24 进行分类，分别得到了 0-9 数字语言识别的三个混淆矩阵，如图 1-图 3 所示。

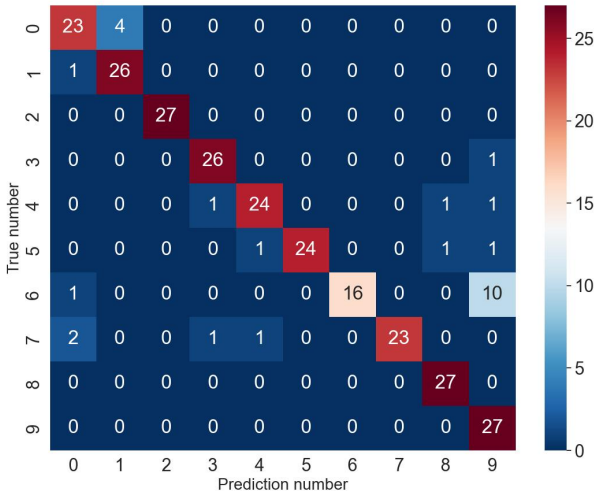

图 1. mfcc 维度为 13 时的分类混淆矩阵

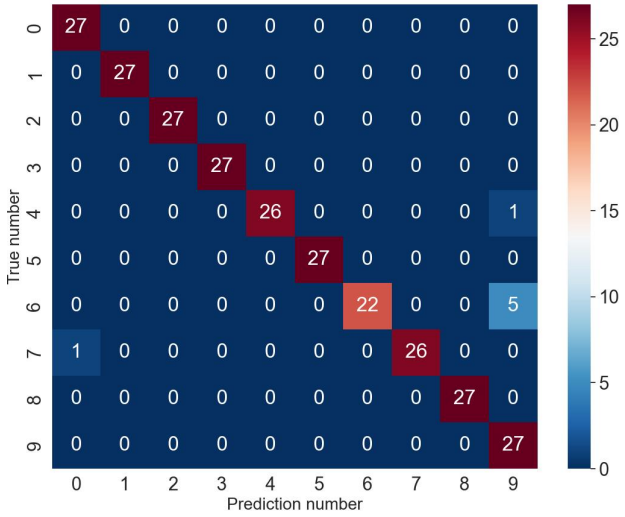

图 2. mfcc 维度为 20 时的分类混淆矩阵

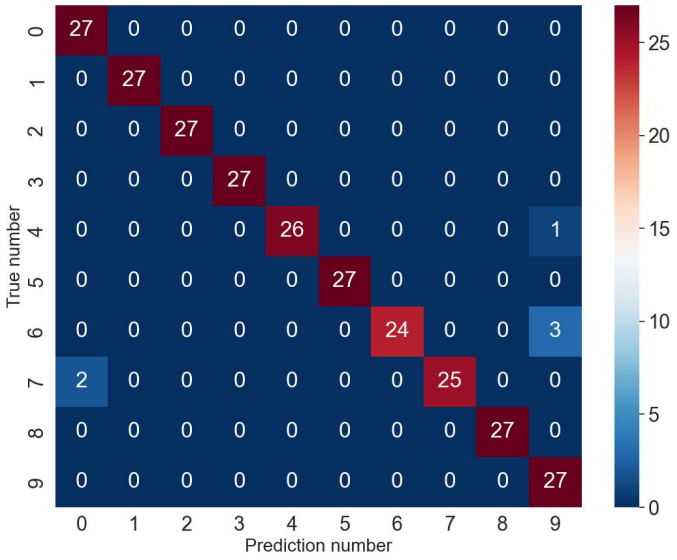

图 3. mfcc 维度为 24 时的分类混淆矩阵

为了更好的比较 mfcc 维度对分类结果的影响，我们对上述三个混淆矩阵进行了更加深入的研究。我们选择了 0-9 中的三个数字，计算出每个混淆矩阵的评价指标，如图 4-图 8 所示。

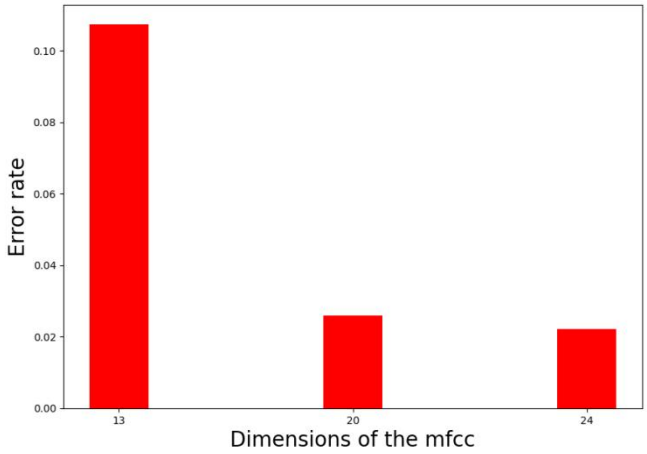

图 4. 三种 mfcc 维度时下的分类 0-9 总错误率

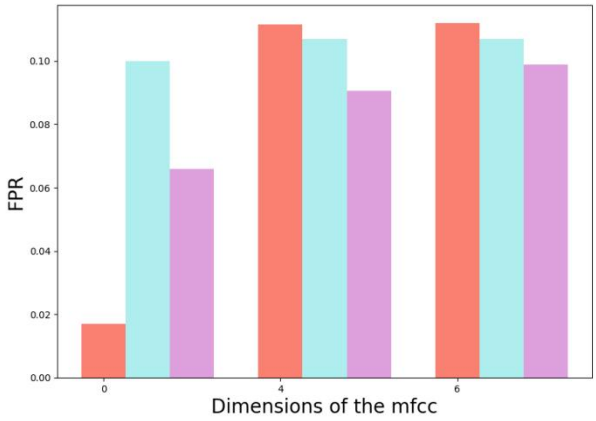

图 5. 三种 mfcc 维度时下的“0”，“4”，“6”的分类 FPR 指标

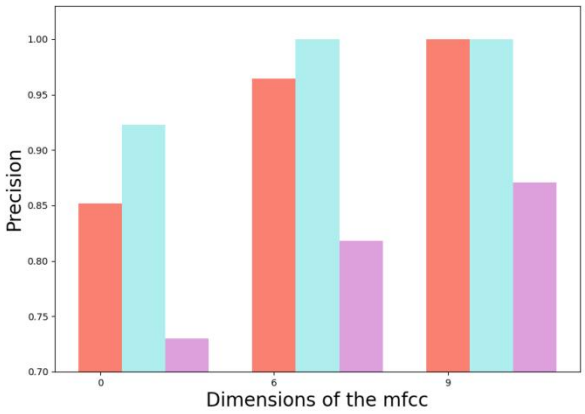

图 6. 三种 mfcc 维度时下的“0”，“6”，“9”的分类 Precision 指标

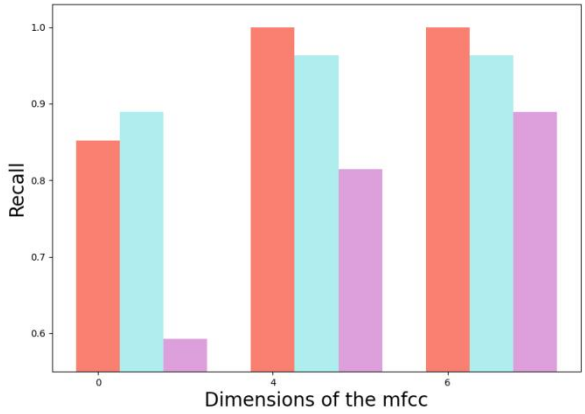

图 7. 三种 mfcc 维度时下的“0”，“4”，“6”的分类 Recall 指标

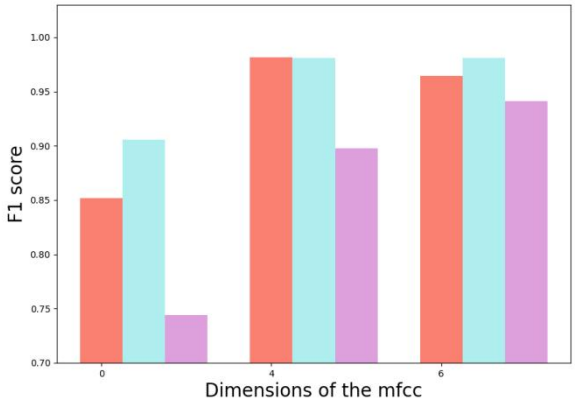

图 8. 三种 mfcc 维度时下的“0”，“4”，“6”的分类 F1 分数

由实验结果可知，mfcc 的特征向量维度为 13 时，分类的结果相较于维度为 20 和 24 的分类结果较差，mfcc 的特征向量维度为 20 的分类结果与维度为 24 的分类结果基本相同。随着 mfcc 特征向量维度的增加，分类的结果越来越好，之后趋于稳定。实验结果表明，适当增加 mfcc 的特征向量维度，有助于提高识别的准确率。

2. 不同的窗函数

我们使用不同的窗函数进行对比实验。我们统一选取 mfcc 特征向量维度为 20，mel 滤波器数量均为 20 的情况，分别使用矩形窗、hamming 窗[3]和 hanning 窗[4]，得到了 0-9 数字语言识别的三个混淆矩阵以及准确率评价指标，如图 9-图 12 所示。

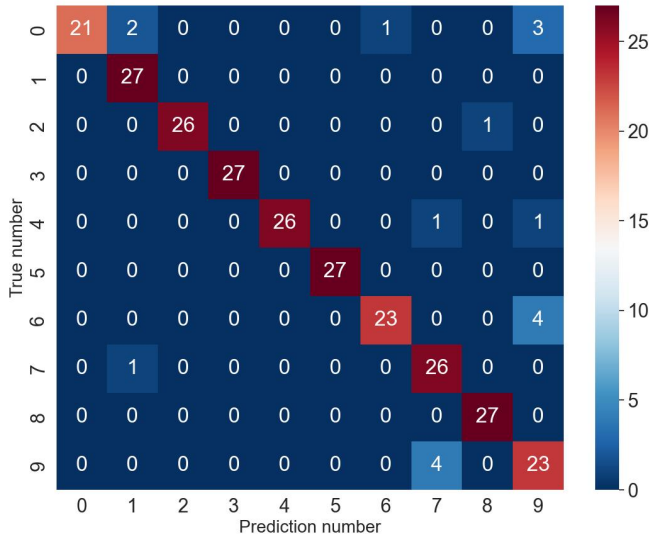

图 9. 使用矩形窗时的分类混淆矩阵

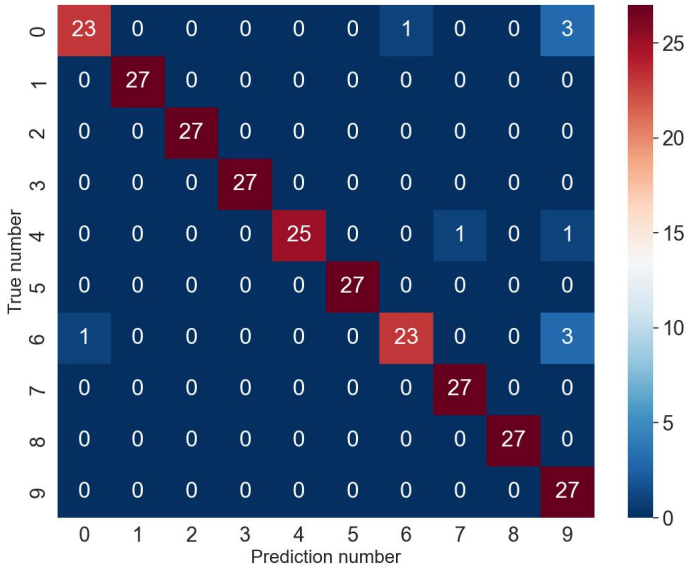

图 10. 使用 hamming 窗时的分类混淆矩阵

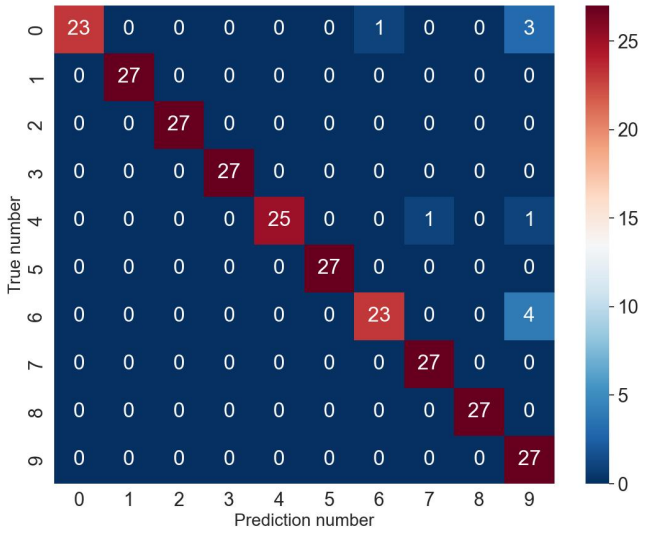

图 11. 使用 hanning 窗时的分类混淆矩阵

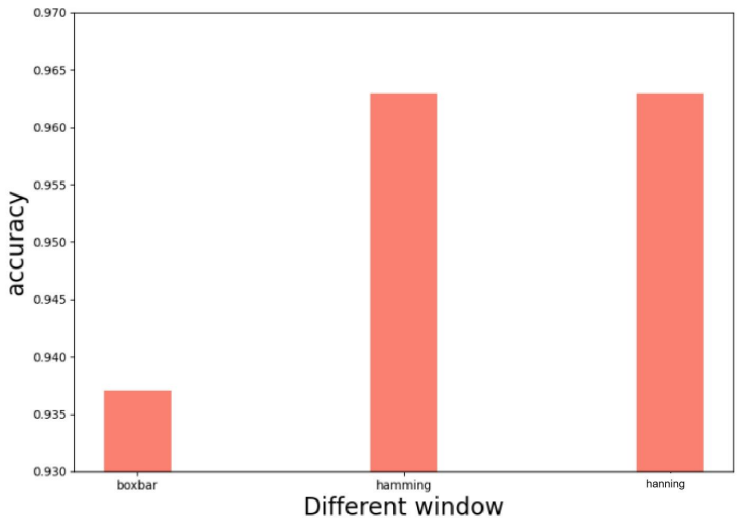

图 12. 使用三种窗函数时的分类准确率指标

由实验结果可知，当窗函数为矩形窗时，分类的结果相较于 hamming 窗和 hanning 窗的分类结果较差，hamming 窗的分类结果与 hanning 窗的分类结果基本相同，但是，三个窗函数的分类结果的准

确率均高于 93%，这说明 mfcc 特征向量维度为 20，mel 滤波器数量为 20 的情况的分类效果较好，为本实验效果较好的参数，并且窗函数的选择在本实验中，对结果的影响属于次要影响因素。

3. 不同的 mel 滤波器组数

我们使用不同的 mel 滤波器组数进行对比实验。我们统一选取 mfcc 特征向量维度为 20，窗函数为矩形窗的情况，分别使用 mel 滤波器组数为 10, 15，20, 40 和 128 三种情况，并且得到了在 20, 40, 和 128 三种情况下，0-9 数字语言识别的三个混淆矩阵以及准确率评价指标，如图 13-图 16 所示。

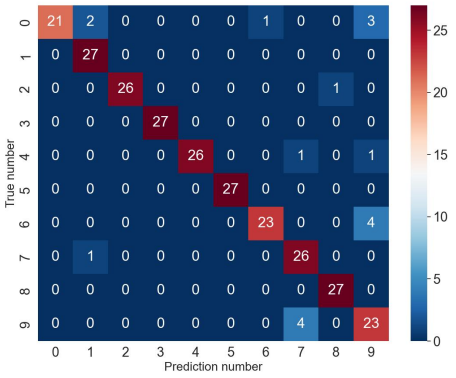

图 13. mel 滤波器组数为 20 时的分类混淆矩阵

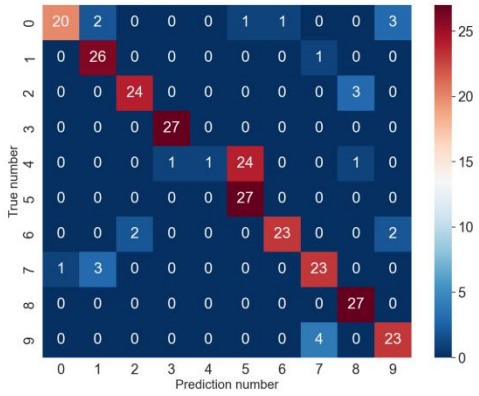

图 14. mel 滤波器组数为 40 时的分类混淆矩阵

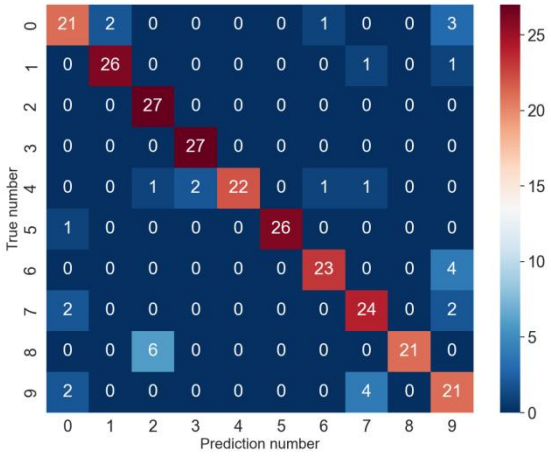

图 15. mel 滤波器组数为 128 时的分类混淆矩阵

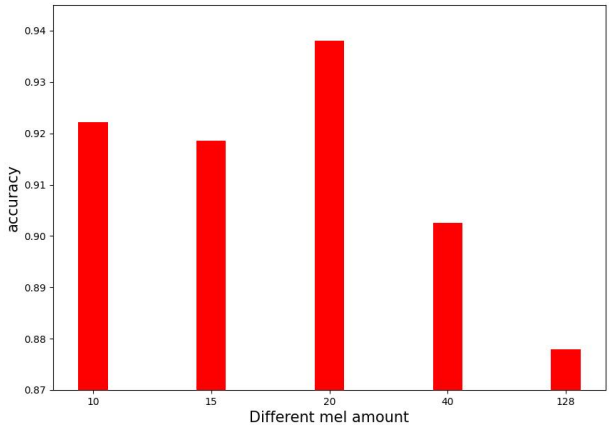

图 16. mel 滤波器组数分别为 10, 15，20, 40, 128 时的分类准确率指标

由实验结果可知，mel 滤波器组数为 128 的分类结果在三者中较差，mel 滤波器组数为 20 的分类结果在三者中较好。实验说明，一味的增加 mel 滤波器组数并不能提高实验结果。mel 滤波器组数为 20，即为本实验较为合适的参数。

GUI 界面

本实验基于 python 平台搭建了上述数字语音识别系统。

我们对模型进行了封装打包，设计了用户友好的 GUI 界面，可以方便快捷地进行 0-9 数字的实时语音识别，如图 17-图 18 所示。

用户既可以选择语音文件进行语音识别，也可以实时录制语音进行识别。我们设计的语音识别系统的实时性强，准确度高，方便用户快捷地进行语音识别工作。

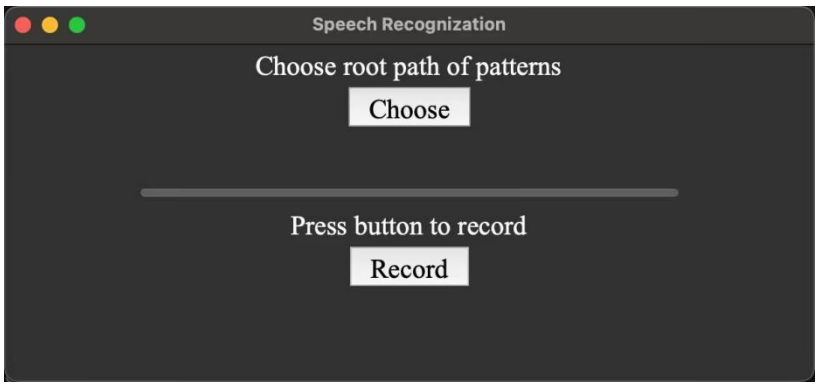

图 17. 语言识别系统的开始界面

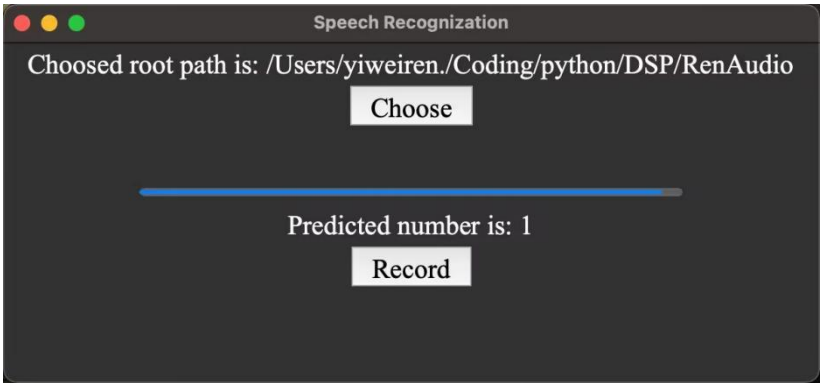

图 18. 语言识别系统的预测界面

结论

本文研究了基于 DTW 和 mel 算法的 0-9 数字信号语音识别方法，通过最优时间扭曲实现语音样本的最佳匹配，构建了 0-9 的数字识别系统，并且实现了用户友好的语音识别 GUI 界面，用户可以方便快捷地进行实时语音识别项目。实验结果表明，我们的方法识别性能好，语音识别实时性强。不仅如此，我们还对模型进行了详细的对比实验与消融实验，并且对实验结果进行了丰富的可视化展示。我们的后续工作将扩充模板规模，探索与深度学习技术的结合，提高识别的复杂度和鲁棒性。

参考文献

1. Zheng, F., Zhang, G., and Song, Z. (2001). Comparison of different implementations of MFCC. J. Comput. Sci. Technol. *16*, 582–589. 10.1007/BF02943243.

2. Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral

Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Preprint at arXiv, 10.48550/arXiv.1003.4083 10.48550/arXiv.1003.4083.

3. Bojkovic, Z.S., Bakmaz, B.M., and Bakmaz, M.R. (2017). Hamming Window to the Digital World. Proc. IEEE *105*, 1185–1190. 10.1109/JPROC.2017.2697118.

4. Fu, W. (2020). Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study. JMIR Med. Inform. *8*, e18677. 10.2196/18677.