

文献引用格式: 郭佳淇, 张继通. 基于 MFCC 和 HMM 的语音识别优化方法研究 [J]. 电声技术, 2024, 48(10): 83–85.

GUO J Q, ZAHNG J T. Research on speech recognition optimization method based on MFCC and HMM [J]. Audio Engineering, 2024, 48(10): 83–85.

中图分类号: TN912

文献标识码: A

DOI: 10.16311/j.audioe.2024.10.024

基于 MFCC 和 HMM 的语音识别优化方法研究

郭佳淇, 张继通

(郑州工业应用技术学院, 河南 郑州 451100)

摘要: 为探究基于梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCC) 和隐马尔可夫模型 (Hidden Markov Model, HMM) 的语音识别优化方法, 首先探讨语音识别系统的基本框架设计, 其次分析 MFCC 特征提取方法, 再次引入期望最大化 (Expectation Maximization, EM) 算法优化 HMM 参数, 最后利用 THCHS-30 数据集进行实验验证。结果表明, 引入 EM 算法优化 HMM, 可有效克服传统 HMM 在复杂语音环境下的识别困难问题, 显著提升系统的识别精度和健壮性。

关键词: 语音识别; 梅尔频率倒谱系数 (MFCC); 隐马尔可夫模型 (HMM); 期望最大化 (EM)

Research on Speech Recognition Optimization Method Based on MFCC and HMM

GUO Jiaqi, ZHANG Jitong

(Zhengzhou University of Industrial Technology, Zhengzhou 451100, China)

Abstract: In order to explore the speech recognition optimization method based on Mel-Frequency Cepstral Coefficients (MFCC) and Hidden Markov Model (HMM), the basic framework design of the speech recognition system is first discussed. Secondly, the MFCC feature extraction method is analyzed, and the Expectation Maximization (EM) algorithm is introduced again to optimize HMM parameters. Finally, the THCHS-30 dataset is used for experimental verification. The results show that the introduction of EM algorithm to optimize HMM can effectively overcome the recognition difficulties of traditional HMM model in complex speech environment, and significantly improve the recognition accuracy and robustness of the system.

Keywords: speech recognition; Mel-Frequency Cepstral Coefficients (MFCC); Hidden Markov Model (HMM); Expectation Maximization (EM)

0 引言

在现代信息社会中, 语音识别技术已广泛应用于智能手机、智能家居等各类智能设备^[1-3]。语音识别系统的核心在于将音频信号转换为对应的文本信息, 从而实现人机交互。随着计算机技术和相关理论不断发展, 语音识别技术取得了显著进步, 但仍面临诸多应用挑战。

目前, 在语音识别任务中, 常见的特征提取方法有梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCC)^[4-5]、线性预测倒谱系数 (Linear Prediction Cepstral Coefficients, LPCC)^[6-7] 等。提

取特征之后, 可以通过隐马尔可夫模型 (Hidden Markov Model, HMM)^[8] 进行语音识别。该方法通过状态转移概率和观测概率的建模, 实现对语音信号的动态时间规整。然而, 传统 HMM 在模型训练和参数估计上存在局限, 难以充分捕捉语音信号的动态特性和多样性。为解决这些问题, 文章提出一种基于 MFCC 和 HMM 的语音识别优化方法。首先, 研究设计语音识别系统的总体框架, 并探讨 MFCC 特征提取方法; 其次, 针对传统 HMM 的不足, 引入期望最大化 (Expectation Maximization, EM) 算法^[9] 优化进行; 最后, 利用 THCHS-30 数据集^[10] 测试和验证提出的方法。

作者简介: 郭佳淇 (2003—), 男, 本科, 研究方向为人工智能。

1 语音识别系统总体框架

为构建一个高效的语音识别系统,文章提出一个基于 MFCC 特征提取和 HMM 优化的语音识别框架。该框架主要包括预处理、特征提取、模型训练与优化以及识别与后处理等多个关键部分,如图 1 所示。

在语音信号预处理阶段,首先对原始语音信号进行降噪处理,去除背景噪声和不必要的干扰。在特征提取阶段,采用 MFCC 方法分析预处理后的语音信号。在模型训练与优化阶段,使用 HMM 对提取的 MFCC 特征进行建模。为提高模型的训练效果和识别精度,引入 EM 算法优化 HMM。在识别与后处理阶段,利用训练好的 HMM 识别输入的语音信号,输出相应的文本信息,并需要对识别结果进行后处理,以提高系统的整体性能。



图 1 语音识别系统框架

2 MFCC 特征提取方法

MFCC 是语音信号处理中广泛使用的特征提取方法,其主要过程包括预加重、分帧与加窗、快速傅里叶变换 (Fast Fourier Transform, FFT)、梅尔滤波器组、对数压缩以及离散余弦变换 (Discrete Cosine Transform, DCT) 等。

预加重的目的是提高高频部分的能量,表达式为

$$y[n]=x[n]-\alpha x[n] \quad (1)$$

式中: $x[n]$ 为输入信号, $y[n]$ 为处理后信号, α 为 0.97, n 为信号样本点的序号。将预加重后的语音信号分成若干帧,并对每帧加窗以减少频谱泄露。使用汉明窗作为窗函数,表达式为

$$w[n]=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

式中: $w[n]$ 为汉明窗函数, N 为窗口长度 (即每帧的样本数)。采用 FFT 将每一帧时域信号转换至频域,表达式为

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi kn}{N}} \quad (3)$$

式中: $X[k]$ 为频域信号, k 为频率索引。将频域信号通过一组梅尔滤波器。使用梅尔滤波器组处理频域信号,可以得到梅尔频谱系数,表达式为

$$S[m]=\sum_{k=0}^{N-1}|X[k]|^2H_m[k] \quad (4)$$

式中: $S[m]$ 为梅尔滤波器的输出, $H_m[k]$ 为第 m 个梅尔滤波器的频率响应。对数运算梅尔滤波器的输出,压缩动态范围为

$$\hat{S}[m]=\log S[m] \quad (5)$$

最后,进行 DCT,得到 MFCC 特征系数,表达式为

$$o[n]=\sum_{m=0}^{M-1}\hat{S}[m]\cos\left[\frac{\pi n(2m+1)}{2M}\right] \quad (6)$$

式中: $o[n]$ 为第 n 个 MFCC 特征系数, M 为梅尔滤波器的数量, m 为累加操作时梅尔滤波器的序号。将语音信号转换成一系列稳定且独特的 MFCC 特征,可为语音识别提供良好的输入信息。

3 引入 EM 的 HMM 优化方法

在获得 MFCC 特征系数后,采用基于 EM 算法优化的 HMM 进行语音识别,可显著改善模型精度和健壮性。

HMM 是一种用于建模时间序列数据的统计模型,通常表示为

$$\lambda=(\gamma, A, B) \quad (7)$$

式中: λ 为 HMM 参数的集合; $\gamma=\{\gamma_i\}$, 为初始状态概率分布,表示系统在初始时刻处于状态 i 的概率; $A=\{a_{ij}\}$, 为状态转移概率矩阵,表示从状态 i 转移到状态 j 的概率; $B=\{b_j(o_t)\}$, 为观测概率分布,表示当状态处于 j 下时, t 时刻观测到的 MFCC 特征系数 o_t 的观测概率。

EM 算法是一种迭代优化算法,通过反复执行期望步骤 (E 步骤) 和最大化步骤 (M 步骤) 估计模型参数,使模型的对数似然函数达到最大化。随机初始化 HMM 的参数 $\lambda=(\gamma, A, B)$ 。在 E 步骤中,在给定当前模型参数 $\lambda^{(k)}$ 的情况下,计算前向概率 $\alpha_t(i)$ 和后向概率 $\beta_t(i)$,表达式为

$$\alpha_t(i)=P(o_1, o_2, \dots, o_t, q_t=S_i | \lambda) \quad (8)$$

$$\beta_t(i)=P(o_{t+1}, o_{t+2}, \dots, o_T | q_t=S_i, \lambda) \quad (9)$$

式中: q_t 为 t 时刻系统处于状态 S_i , T 为观测序列的长度。计算观测序列在 t 时刻从状态 S_i 转移到状态 S_j 的概率,表达式为

$$\xi_t(i, j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O | \lambda)} \quad (10)$$

式中： $P(O|\lambda)$ 为给定模型参数下观测序列的概率，可以通过前向算法或后向算法计算得到； $b_j(o_t)$ 为当状态处于 j 下时， t 时刻观测到的MFCC特征系数 o_t 的观测概率。计算观测序列在 t 时刻处于状态 S_i 的概率，表达式为

$$\gamma_i(i) = \sum_{j=1}^N \xi_i(i, j) \quad (11)$$

在 M 步骤中，根据E步骤计算得到的期望值，更新模型参数，表达式为

$$\gamma_i^{(k+1)} = \gamma_i(i) \quad (12)$$

$$a_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \xi_i(i, j)}{\sum_{t=1}^{T-1} \gamma_i(i)} \quad (13)$$

$$b_j^{(k+1)}(o_t) = \frac{\sum_{t=1}^T \gamma_i(j) \delta(o_t, v_k)}{\sum_{t=1}^T \gamma_i(j)} \quad (14)$$

式中： $\delta(o_t, v_k)$ 为指示函数，当 $o_t=v_k$ 时取1，否则取0； k 为当前的迭代次数。

迭代执行EM算法的E和M两个主要步骤，直至模型参数趋于稳定，即对数似然函数的变化值低于设定阈值。在实际应用中，利用优化后的HMM评估给定观测序列的概率，并推断最佳状态序列，从而完成语音到文本的转换过程。

4 实验与分析

在MATLAB平台，使用THCHS-30数据集测试根据文章方法优化后的HMM。首先，使用MATLAB的Signal Processing Toolbox预处理语音信号，包括降噪、分帧、加窗，帧长设为25 ms，帧移为10 ms，加窗采用汉明窗。其次，提取信号的MFCC特征向量。具体实现中，可使用Voicebox工具箱中的mfcc函数，MFCC系数为13个。再次，构建HMM，初始化初始状态概率、状态转移概率、观测概率等模型参数，并通过EM算法优化HMM，迭代次数设为100次或直至对数似然函数收敛，收敛阈值设为 10^{-4} 。最后，利用优化后的HMM对测试集进行语音识别。通过计算识别精确率、准确率、召回率评估模型性能，结果如表1所示。

精确率反映了系统在识别结果中的准确性，准确率反映系统对所有样本的整体识别能力，召回率量化系统检测出正样本的比例。从表1可以看出，

优化HMM的精确率、准确率、召回率分别达到0.92、0.89、0.91，明显高于标准HMM的0.85、0.82、0.87。综合来看，优化后的HMM能够有效降低误识别率，正确识别语音输入，与标准HMM相比性能略有提升。

表1 标准HMM与优化HMM对比结果

方法	精确率	准确率	召回率
标准HMM	0.85	0.82	0.87
优化HMM	0.92	0.89	0.91

5 结语

文章在探讨语音识别技术的同时，深入分析MFCC特征提取方法及其在语音信号预处理中的作用。通过引入EM算法优化HMM，有效克服传统HMM在复杂语音环境下的识别困难问题，显著提升系统的识别精度和健壮性。实验结果验证了文章方法的有效性，证明其在现有语音识别技术中具有创新和实用价值。未来的研究方向可以进一步探索如何结合深度学习方法优化语音识别系统，以应对更加复杂和多样化的语音数据，进一步提升识别性能和应用广泛性。

参考文献：

- [1] 应长鸣, 何志学. 利用语音识别实现盲人对智能手机的声控操作[J]. 电脑知识与技术, 2022, 18(9): 49-51.
- [2] 黄玲, 周裕滨, 黄源俊. 基于语音识别技术的智能家居系统设计[J]. 电脑知识与技术, 2023, 19(31): 38-40.
- [3] 赵一鸣, 陈宇, 刘齐平. 智能语音助手用户研究: 理论进展与实践启示[J]. 数字图书馆论坛, 2023, 19(5): 26-34.
- [4] 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法[J]. 计算机应用, 2012, 32(9): 2542-2544.
- [5] 李志忠, 滕光辉. 基于改进MFCC的家禽发声特征提取方法[J]. 农业工程学报, 2008(11): 202-205.
- [6] 解滔, 郑晓东, 张葵. 基于线性预测倒谱系数的地震相分析[J]. 地球物理学报, 2016, 59(11): 4266-4277.
- [7] 艾长胜, 何光伟, 董全成, 等. 基于切削声LPCC的刀具磨损监测[J]. 中国机械工程, 2009, 20(17): 2045-2048.
- [8] 张经, 杨健, 苏鹏. 语音识别中单音节识别研究综述[J]. 计算机科学, 2020, 47(增刊2): 172-174.
- [9] 刘铭, 于子奇. 一种改进的期望最大化算法[J]. 吉林大学学报(理学版), 2022, 60(5): 1176-1182.
- [10] 李荪, 曹峰, 刘姿杉. 面向算法模型的语音数据集质量评估方法研究[J]. 计算机科学, 2022, 49(增刊2): 519-524.

编辑: 李如冰