对 MFCC 进行 GMM 聚类的汉语数字识别方法①

高文曦，于凤芹

(江南大学 物联网工程学院，无锡 214122)

摘 要: 汉语数字识别常用 MFCC 作为特征，针对 0-9 十个数字 MFCC 样本特征数据量大的问题，提出了用 GMM 模型对提取的特征参数 MFCC 的数据进行聚类来减少数据量，以 GMM 模型参数中的均值作为新的特征，采用动态规划算法进行汉语数字语音识别。仿真实验表明，进行 GMM 特征变换后的新特征数据为 MFCC 的 30.9%，系统运行时间减少了 237.18s，识别率降低 1.11%。

关键词: 汉语数字识别；MFCC；GMM 聚类

Chinese Digital Identification Based on MFCC by Using GMM Clustering

GAO Wen-Xi, YU Feng-Qin

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: MFCC is widely used in Chinese digital identification. Because the amount of MFCC extracted from 0-9 is too large, the mean of model parameters which is clustered with GMM by MFCC to reduce the amount is employed as a new feature with DTW for Chinese digital identification. Simulation results demonstrate that the amount of the new feature is 30.9% to that of MFCC, the running time reduces by 237.18s, but the recognition rate decreases by 1.11%.

Key words: Chinese digital identification; MFCC; GMM clustering

1 引言

汉语数字语音识别的任务是识别“0”到“9”等十个非特定人汉语数字语音[1]。目前在汉语数字识别中常用的特征参数有：线性预测倒谱系数（Linear Predictive Cepstral Coefficient, LPCC）和 Mel 倒谱系数（Mel Frequency Cepstrum Coefficient,MFCC）。谱参数 MFCC 是基于 Mel 尺度模拟人耳的听觉特性，具有较好的区分能力，因此，得到了广泛的应用，文献[2]采用 MFCC 与 LPCC 相结合，进行英文单词识别，平均识别率为 91.65%。文献[3]采用分数傅里叶变换，将 MFCC 推广到分数形式，应用到说话人识别系统，识别率达到 93%。文献[4]对 MFCC 参数提取模块提出新的提取算法，相对于标准算法，其在特征提取模块部分减少了 53%的计算量，但是识别率降低了 1.7%。文献[5]采用 MFCC 与 PLP 相结合，分别在纯净语音和和含噪环境下，得出其具有很好的抗噪性。不同参数的组合可以更加准确的代表样本特征或具有更好的抗噪性能，但是所需处理的数据增多，识别时间长。

本文采用特征参数 MFCC 作为 0-9 十个数字的特征参数，在提取过程中发现样本特征参数的数据量较大，增加了识别模块的复杂度。高斯混合模型（Gaussian Mixture Model, GMM）是一种典型的生成式模型，能够快速有效地处理大量训练数据，本文采用 GMM 模型对提取的特征参数的数据进行聚类，以高斯模型的均值参数作为新的特征，采用动态时间规整（Dynamic Time Warping, DTW）算法进行汉语数字语音识别。仿真实验表明，经过 GMM 聚类变换后得到的均值特征参数，减少了识别模块训练的数据量，提高了识别速度。

2 对MFCC进行GMM聚类的汉语数字识别原理

2.1 基本流程

对 MFCC 进行 GMM 模型聚类的汉语数字识别的流程为：首先读取 0—9 十个数字的语音，然后分别提

① 基金项目:国家自然科学基金(61075008)
收稿时间:2011-03-15;收到修改稿时间:2011-04-13

取每个数字的特征参数 MFCC，采用 GMM 模型对特征参数 MFCC 的数据进行聚类，以高斯模型参数中的均值矢量作为新的特征，采用动态时间规整算法进行分类，输出匹配模板的序号，统计正确个数计算识别率。对 MFCC 进行 GMM 模型聚类的汉语数字识别的原理框图如图 1 所示：

图 1　对 MFCC 进行 GMM 聚类的汉语数字识别的原理图

2.2 MFCC 基本原理

人耳对不同频率的语音信号具有不同的感知能力，在 1000Hz 以下，感知能力与频率成线性关系，而在 1000Hz 以上，感知能力与频率成对数关系。因此人们提出了 *Mel* 频率的概念，其意义为：1 *Mel* 为 1000Hz 的音调感知程度的 1/1000。频率 f 与 *Mel* 频率之间的转换公式为：

$$Mel(f)=2595\lg(1+f/700) \tag{1}$$

式中，f 为频率。

Mel 滤波器组的频带划分着眼于人耳听觉机理，将语音的实际频率变换到感知频率中，能更好地模拟人的听觉处理过程。每个 *Mel* 三角带通滤波器的传递函数：

$$\mathrm{H}_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)<k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k\le f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{2}$$

其中，$0<m\le M$，M 为滤波器的个数，$f(m)$ 为每个三角滤波器的中心频率。将经过 *Mel* 频率取对数能量，再将对数能量做离散余弦变换，得到 MFCC 特征参数。

采用传统的 12 阶 MFCC 作为汉语数字语音的特征参数，由仿真实验统计其平均每个数字的样本特征数据量为 738.6 个数据。

2.3 基于 EM 算法的 GMM 聚类模型

GMM 模型是一种典型的生成式模型，能够快速有效地处理大量训练数据，其训练过程是对特征参数集的聚类过程。高斯混合模型就是采用这 M 个单高斯分布的线性组合来描述语音信号在特征空间的分布：

$$P(X_t\mid\lambda)=\sum_{i=1}^{M}w_ib_i(X_t) \tag{3}$$

$$b_i(X_t)=P(X_t\mid i,\lambda)=\frac{1}{(2\pi)^{K/2}|\Sigma_i|^{1/2}}\times\exp\left\{-\frac{1}{2}(X_t-\mu_i)'\sum\nolimits_i^{-1}(X_t-\mu_i)\right\} \tag{4}$$

其中，X_t 是长度为 $\mathrm{T}\{t=1,...,T\}$ 的 K 维语音特征矢量，M 是高斯分量个数；$b_i(X_t)$ 是均值为 μ_i，协方差矩阵 Σ_i 的高斯函数；w_i 为其权重，有 $\sum_{i=1}^{M}w_i=1$。整个高斯混合模型参数集用 λ 来描述：$\lambda=\{w_i,\mu_i,\sum_i\},i=1,...,M$。在高斯混合模型训练过程中，在极大似然准则下，利用最大期望算法（Expectation Maximization, EM）进行参数 λ 估计，先给 λ 赋一个初值，然后估计新参数 λ'，新参数再作为当前参数进行训练，不断迭代，各参数的重估公式为：

$$w_i=\frac{1}{T}\sum_{t=1}^{T}P(i/X_t,\lambda) \tag{5}$$

$$\mu_i=\frac{\sum_{t=1}^{T}P(i/X_t,\lambda)X_t}{\sum_{t=1}^{T}P(i/X_t,\lambda)} \tag{6}$$

$$\sigma_i^2=\frac{\sum_{t=1}^{T}P(i/X_t,\lambda)(X_t-\mu_i)^2}{\sum_{t=1}^{T}P(i/X_t,\lambda)} \tag{7}$$

将特征参数 MFCC 经过 GMM 模型进行聚类变换，以 GMM 模型的均值参数作为新的参数，由仿真实验可知，其平均每个数字的样本特征数据量为 228 个数据。

2.4 DTW 识别器

动态时间规整技术就是把时间规整和距离测度计算结合起来的一种非线性规整技术，它成功的解决了发音长短不一的范本匹配问题[6]。DTW 算法原理如下：假设测试语音参数共有 I 帧向量，而参考范本共有 J 帧向量，且 $I\neq J$，则动态时间规整就是要寻找一个时间规整函数 $j=w(i)$，它将测试向量的时间轴 i 非线性地映像到模板的时间轴 j 上，并使函数满足：

$$D=\min\sum_{i=1}^{I}d[T(i),R(w(i))] \tag{8}$$

式中，$d[T(i),R(w(i))]$ 是第 i 帧测试向量 $T(i)$ 和第 j 帧范本向量 $R(j)$ 之间的距离测度，D 则是处于最优时间规整情况下两向量的距离。

3 基于GMM模型聚类的特征变换

基于 GMM 模型聚类特征变换流程图如图 2 所示：

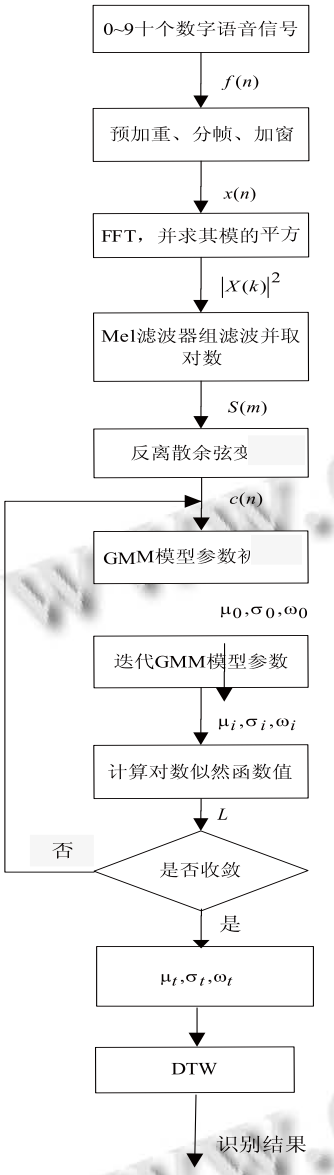

图 2 对 MFCC 进行 GMM 模型聚类的汉语数字识别流程图

基于 GMM 模型聚类的特征变换具体步骤如下：

1) 读取 0-9 十个数字语音信号 *f(n)* 经过预加重分帧、加窗等处理，得到每个语音帧的时域信号 $x(n)$；

2) 将时域信号 $x(n)$ 经过 FFT 模块后得到线性频谱 $X(\mathrm{k})$，并求其模的平方 $|X(\mathrm{k})|^2$；

3) 将上述线性频谱通过 Mel 滤波器组得到 Mel 频谱，并通过对数能量的处理得到对数频谱，$S(m)=\ln\left[\sum_{k=0}^{N-1}|X(k)|^2 H_m(k)\right], 0\le m\le M$，$H_m(k)$ 为 Mel 滤波器的传递函数，每个滤波器具有三角形滤波特性；

4) 将上述对数频谱经过反离散余弦变换 $c(n)=\sum_{m=0}^{M-1}S(m)\cos\left[\frac{\pi n(m+1/2)}{M}\right], 0\le m\le M$，得到 MFCC 特征参数；

5) 用 k 均值聚类算法选取 GMM 模型的参数的初始值。本仿真实验所用 k 均值聚类函数为 EM 算法中自带的函数 kmeans.m。输入为 MFCC 特征参数，输出为高斯模型参数的初始值；

6) 将原模型参数值带入迭代公式(5)、(6)、(7)更新模型参数，并计算新的模型参数下的似然函数值；

7) 计算对数似然函数值。似然函数 $L(X\mid\lambda)=\prod_{i=1}^{T}f(X_i\mid\lambda)$，其中 $f(X_i\mid\lambda)$ 为高斯模型的概率密度函数，$X=(c(1),c(2),...,c(n))^T$；

8)收敛条件：如果 $100*\frac{\log L(X\mid\lambda_t)-\log L(X\mid\lambda_{t-1})}{\log L(X\mid\lambda_{t-1})}<\varepsilon$，即认为收敛，其中 ε 取 0.1；

9) 如果收敛，输出模型参数 λ_t；如不收敛，重复步骤 6)7)8)，直至收敛为止或达到最大迭代次数 1000 次；

10) 将模型的均值参数输进 DTW 识别器，输出识别结果。

上述过程可以把 GMM 模型中的每一个高斯分布看作一类，对 MFCC 的特征参数集进行聚类，其中均值参数反映了不同样本特征在特征空间的相对位置，是 GMM 模型中的一个重要参数，方差反映数据分布的密集程度，权值表示该类数据的多少。将 GMM 模型的均值矢量组合在一起构成新的特征参数，与原有的特征参数相比，减少了数据。

4 仿真实验及其结果分析

仿真实验中所用的语音库由 10 人录制，每人每个数字发音 2 遍，共 200 个样本，其中 100 个样本用于训练，100 个样本用于测试。采用 22050Hz 的采样频率，单声道，16 位采样精度进行语音信号的录制。

4.1 数据量的比较

对 100 个用于训练的语音样本提取其参数 MFCC，将 MFCC 的特征矩阵送入 GMM 模型得到 GMM 模型的均值参数，以平均每个样本特征的数据量进行对比分析，统计结果如表 1 所示：

表 1 MFCC 与 GMM 模型的均值参数的数据处理量

参数	MFCC	GMM 模型的均值参数
平均处理数据量/字	738.6	228

由上表可以看出，平均每个数字 GMM 模型均值参数的数据量是 MFCC 的 30.9%，由此可以得出采用 GMM 模型对参数 MFCC 特征数据进行特征转换得到的均值参数和特征参数 MFCC 相比减少了输入识别模块的数据量，提高了系统的性能。

4.2 识别率和运行时间的比较

将剩余的 100 任取 90 个作为测试样本，采用 DTW 识别器进行识别，以与测试模板距离最小的模板序号作为识别结果进行输出，比较 MFCC 和 GMM 模型的均值参数的识别效果，统计不同特征参数的平均识别率和系统运行时间，其结果如表 2 所示：

表 2 MFCC 与 GMM 模型均值参数识别率及识别时间的比较

参数	平均识别率	运行时间
MFCC	90%	287.04s
GMM 模型的均值参数	88.89%	49.86s

由上表可知，采用 GMM 模型的均值参数其系统的平均识别率和 MFCC 相比下降了 1.11%，但是时间缩短了 237.18s，由此可见，采用 GMM 模型对特征参数进行聚类变换，减少了输入识别器的数据，提高了识别的速度。

5 结论

本文采用 GMM 模型对样本特征参数 MFCC 进行聚类，用 GMM 模型的均值参数作为新的特征参数，其均值参数的数据量相比于特征参数 MFCC 的数据量减少了 510 个，识别时间缩短了 237.18s，但识别率只降低了 1.11%，有效地提高了识别速度，验证了该方法的有效性。但是在语音中加入高斯白噪声，信噪比为 10dB，识别率只有 53.33%，可知经过 GMM 模型进行聚类得到的均值参数易受噪声影响，其特征参数的鲁棒性是今后研究的一个方向。

参考文献

1 马静.基于 HMM 模型的汉语数字语音识别算法的研究.太原:太原理工大学,2008.
2 李萱.语音特征参数提取方法研究.西安:西安电子科技大学,2006.
3 张永亮,张先庭,鲁宇明.基于 FMFCC 和 HMM 的说话人识别.计算机仿真,2010,27(5):352–357.
4 张晶,范明,冯文全.基于 MFCC 参数的说话人特征提取算法的改进.电声技术,2009,33(9):61–66.
5 Li Q, Huang Y. Robust speaker identification using an auditory-based feature.In:2010 IEEE International Conference on Acoustics Speech and Signal Processing. Dallas, 2010,4514–4518.
6 赵力.语音信号处理.北京:机械工业出版社,2003.

(上接第 173 页)

4 Paute J, Jordan F. Using Fractal compression scheme to embed a digital signature into an image. Proc SPIEPhotonics East Symposium,Boston,USA,1996.
5 Pi MH, Li CH, Li H. Anovel fractal image watermarking watermarking. IEEE Trans. on Multimedia, 2006,8(3): 488–499.
6 Xie RS, Yang SG. A Digital Image Water marking Method Based on Fractal Transform in DWT Domain. 1st International Conference on Modelling and Simulation, Nanjing, China, 2008.
7 陈辉,郭科,陈聆.基于分型编码的遥感影像数字水印技术研究.计算机应用研究,2007,24(9):83–85.
8 冯茂岩,冯波,沈春林.基于分块 DCT 变换和 Arnold 置乱的自适应图像水印算法.计算机应用,2008,28(1):171–173.
9 韩强,马洪.DCT 域上基于 HVS 的盲水印添加方法.四川大学学报(自然科学版),2005,42(3):444–449.
10 刘方,杨峰.改进的基于 DCT 的加密盲水印算法.计算机工程与应用,2009,45(13):124–126.