一种基于 HTK 的数字语音识别系统①

魏 巍，张海涛

(辽宁工程技术大学 电子与信息工程学院，葫芦岛 125105)

摘 要：数字语音识别是语音识别一个极其重要的分支，其在现实生活中的应用愈加广泛。HTK 是英国剑桥大学开发的一套基于 C 语言的语音处理工具箱，广泛应用于语音识别、语音合成、字符识别和 DNA 排序等领域。从 HTK 的基本原理和软件结构出发，设计了一个基于 HTK 的数字语音识别系统，并验证了其识别效率。随后，通过更换识别单元，更改特征参数的维数和增加高斯混合分量的个数来考虑不同因素对系统性能的影响。最后，通过比较试验，验证了识别单元、高斯混合分量的数目以及 MFCC 维数的适当组合可提高系统的正确识别率。

关键词：语音识别；HTK；HMM；识别单元；MFCC

Digital Speech Recognition System Based on HTK

WEI Wei, ZHANG Hai-Tao

(School of Electronics and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Digital speech recognition is an extremely important branch of speech recognition. Its application in real life is used more and more widely. HTK is a C language-based toolkit developed by CUED mainly used for speech signal reorganization, speech synthesis, character reorganization, DNA compositor and so on. From HTK's general principles and software architecture, this paper designs a digital speech recognition system based on HTK, and verifies its recognition efficiency. Then by changing the identification unit and MFCC dimension, and by increasing the number of gaussian mixture components, it considers effects of different factors on the performance of the system. Finally, through the comparing test, it verifies the right combination of the identification unit and the number of gaussian mixture components, and also proves that MFCC dimension can enhance the system's correct rate.

Key words: speech recognition; HTK; HMM;identification unit; MFCC

1 引言

随着计算机与信息技术的继续发展，语音交互技术必将成为人机交互的必要手段[1]。语音识别技术就是让机器听懂人类的语音并执行相关的动作，是一个研究的热点。语音识别系统根据不同的准则可以分为孤立词、连接词和连续词识别；小词汇量、大词汇量识别；特定人、非特定人的语音识别系统。连续数字语音识别是语音识别的一个重要分支，数字语音识别，尤其是连续数字识别无论在大词表的语音识别系统还是小词表语音识别系统中都具有重要的意义，因此，它在现实中具有广泛的应用前景，在互联网，通信，军事，国防，人机交互等方面都有重要的应用价值。虽然这方面的研究有很多，但目前仍有许多问题有待进一步探索。本文结合隐马尔可夫模型原理，利用 HTK（HMMToolKit）语音处理工具箱，实现了数字语音识别系统的设计。并且从识别单元（音节和声韵母）、更改特征参数的维数和增加高斯混合分量的个数来考虑选取不同因素的情况下对本系统性能的影响。最终，通过实验证明了识别单元、高斯混合分量的数目以及 MFCC 维数的适当组合可提高系统的正确识别率。

隐马尔可夫模型是一种用参数表示的，用于描述随机过程统计特性的概率模型，它是由马尔可夫链演变而来的。如果在分析的区间内信号是非时变或平稳

① 收稿时间:2010-12-16;收到修改稿时间:2011-04-10

的，那么用线性模型就可描述它；但如果在分析的区间内信号是时变的，则线性模型的参数也是时变的。所以，最简单的方法是在极短的时间内用线性模型参数来表示，然后，再将许多线性模型在时间上串接起来，这就是马尔可夫链。由于不能准确地确定信号的时长，所以用马尔可夫链描述时变信号不是最佳和最有效的。而隐马尔可夫模型既解决了短时模型描述平稳段的信号，又解决了每一个短时平稳段是如何转变到下一短时平稳段的问题。它利用概率及统计学理论成功地解决了如何辨识具有不同参数的短时平稳的信号段以及如何跟踪它们之间的转化等问题。由于语言的结构信息是多层次的，除了语音特性外，还牵涉到音长、音调、能量等超音段信息以及语法、句法等高层次语言结构的信息。而隐马尔可夫模型既可描述瞬变的随机过程，又可描述动态的随机过程的转移特性，所以它能够利用这些超音段和语言结果的信息。

1　HTK[2]语音识别的基本原理

HTK（Hidden Morkov Model Toolkit）是英国剑桥大学工程系开发的一套构建隐马尔可夫模型（HMMs）的工具集。其广泛应用于语音识别、语音合成、字符识别和 DNA 排序等领域。但工具集设计的目的主要是建立语音识别系统。语音识别技术主要有特征提取、模型训练以及模式匹配准则三个方面，此外，还涉及到语音识别单元的选取。语音识别系统主要包括语音特征参数的数据准备工具、模型训练工具、语音识别工具和模型分析工具。目前，HTK 的稳定发行版本是 3.4 版本[3]，所有代码使用 c 语言编写，可以在 Windows 和 Linux 上编译使用。HTK 的特点是开放源代码，用户可以在分析其代码的基础上，对某些算法和模块进行修改。HTK 工具包的结构如图 1 所示：

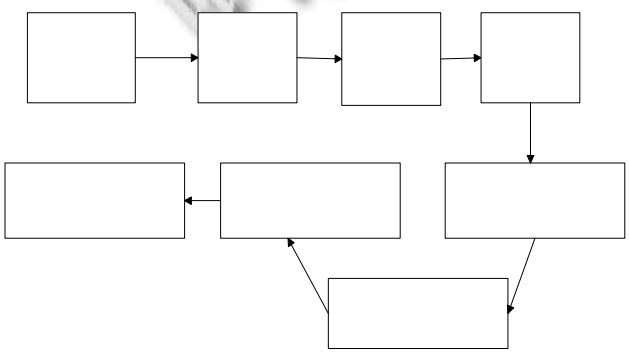

图 1　HTK 工具包结构图

1.1 数据准备工具

为了构造一组 HMMs，需要一系列的语音文件和与其相关的标记文件。在数据准备阶段，要进行语法文件的创建、录音并标记、语音识别单元元素列表文件创建等工作。其中最重要和最费时间的是录音并且标记，可以用 HTK 提供的语音录音程序 HSLab 对声音进行录音，也可以用 Praat、Cooledit 等录音程序进行语音资料的收集。HSLab 工具可以同时被用于录制语音文件和手工标注，可标注为不同级的需要的脚本。HCopy 用于将一个或多个原始的语音波形文件转换成所需的特征参数的输出文件。

1.2 训练工具

训练的目的是让系统从量的真实语音中学习必要的模型参数形成语音参考模式库，为识别阶段做准备。在训练过程中，需要对各个语音识别单元建立所需的 HMM 模型。所用到的工具主要有 HInit 或 HCompv 和 HRest 三个工具，训练工具主要基于 Baum-welch 重估算法，训练时采用嵌入式的方式。在训练过程中,需要对各个语音识别单元建立所需的 HMM 模型，HCompv，HInit 用于估计出一套初始模型参数。HRest 用于重估参数，基于 Baum-Welch 重估算法。

1.3 识别工具

模型识别即是指根据模型训练得到的模型参数对输入的语音进行模式匹配，并给出相应的识别结果。HTK 提供唯一的识别工具 HVite 工具，这是基于 Viterbi 搜索算法的识别工具。在使 HVite 工具的过程中需要两个专用的配置文件：一是所需识别语音的语法网络文件；二是识别语音文件所对应的一本发音词典。其中，利用 HParse 工具可以建立所需的语法网络文件，直接将词典语法写入文本文档即可以建立所需的发音字典文件。最后用 HSGen 对所建立的网络文件和词典语法进行测试。

1.4 分析工具

HTK 提供模型性能分析工具 HResults 工具，用于分析识别率。识别结果的评价主要是通过 HTK 工具中的评价工具 HResults 来实现的，最终给出的结果包括句子和词的识别率以及其他的信息。

2　基于HTK的语音识别系统的搭建

由于数字语音识别在实际生活中具有很广泛的应用，本文设计了一个基于 HTK 的数字语音识别系统[4]。

根据 HTK 语音识别的原理，将该系统的搭建过程分成四步：数据准备，模型训练，模式识别和模型分析。

使用 HTK 构建的数字语音识别系统的体系结构如图 2 所示：

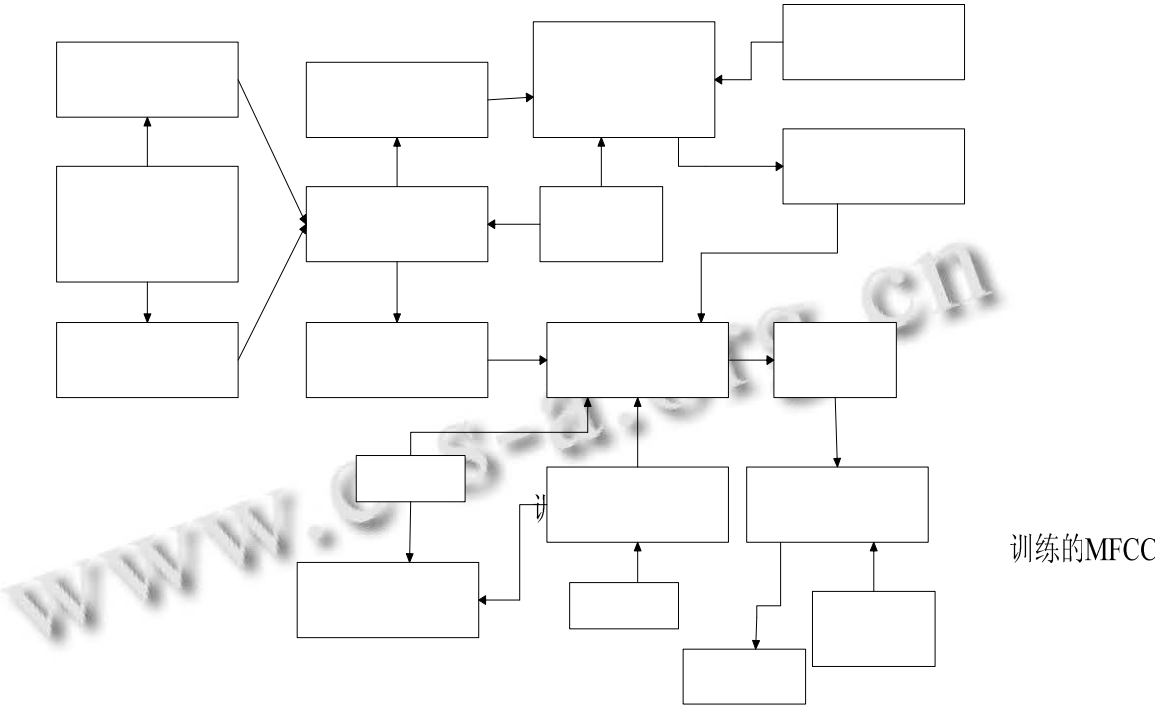

图 2 基于 HTK 构建的数字语音识别系统的体系结构图

2.1 数据准备

在数据准备阶段要准备训取训练语音数据的特征，使用语音录音程序 HSLab 对“0~9”这十个声音进行录音，并对每个录音文件进行标注，对语音文件的标注，可采用手工标注方式，也可以采用自动标注方式，或是自动和手工相结合。这里，采用手工标注的形式。对于每个录音文件，需要标注三个连续的区域：开始停顿（标记为 sil）、录音文件[5]（标记为 0~9 对应的拼音）、结束停顿（标记为 sil）。例如，数字语音“0”的标注如图 3 所示。在完成特征提取之后，随即进行特征参数的提取。在本系统中，所采用的特征参数为 MFCC（Mel Frequency Cepstral Coefficient），MFCC 是 Mel 倒谱系数，其模拟了人的听觉特性，是符合人听觉特性的语音特征参量，在实际应用中可以取得较高的识别率。进行语音特征提取，也就是要将每一个语音文件转换成 MFCC 文件，使用 HCopy 工具即可完成本使命，本系统中对每个信号帧提取 39 维 MFCC_0_D_A（其中，0 表示有倒谱系数，D 表示有一次差分系数，A 表示有二次差分系数）特征参数，为下面进行语音识别奠定基础。

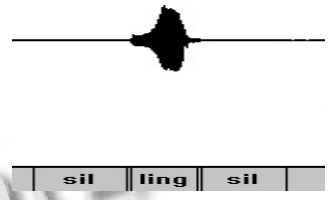

图 3 数字语音“0”的标注图

2.2 模型训练

在训练过程中，系统中每个音子均用一个 HMM 来表示，模型为自左至右的 HMM 结构，对语音而言，由于每个词的时序关系可以通过状态的先后关系来体现，通常都采用自左至右的模型。因此为每个识别单元创建一个 HMM 模型。本系统中，创建具有 5 个转移状态的 HMM 模型，采用的是从左到右的拓扑结构（见图 4）。其中，状态 1 和状态 5 分别为初始状态和终止状态，称为非发散状态；状态 2、3 和 4 称为活动状态。本系统中 HMM 模型对应的状态转移矩阵为 5x5 的转换矩阵，见表 1。在训练过程开始之前，为了使得训练算法快速精准收敛，HMM 模型参数必须根据训练数据正确初始化。将所有特征矢量的均值和协方

差作为模型的初始均值和协方差，随后，通过使用 HTK 提供的 2 套训练工具：HInit/HRest 或 HCompv，即可实现对所以模型的初始化和训练工作。如果训练语料声学单元边界已完成标记，可以使用 HInit / HRest 进行单词分离式训练。一般情况下，调用 HCompv 用全局均值和方差初始化原始模型的高斯参数，接着多次调用 HRest 进行 Baum-Welch 重估。

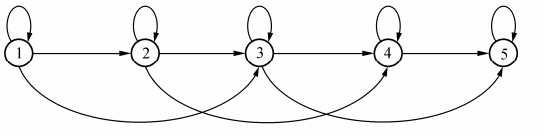

图 4　具有 5 个状态的由左至右的 HMM 模型

表 1　5x5 的状态转移矩阵

0.0	1.0	0.0	0.0	0.0
0.0	0.6	0.4	0.0	0.0
0.0	0.0	0.6	0.4	0.0
0.0	0.0	0.0	0.7	0.3
0.0	0.0	0.0	0.0	0.0

2.3 模式识别

语音识别实质上就是采用搜索算法寻找最有可能的结果。在连续语音识别问题中，所有音子模型的状态按时间重复排列，每一个时间点对应一帧语音特征矢量，形成一个状态矩阵，这样每一帧语音特征矢量就会与所有的 HMM 状态对应起来[6]。HTK 提供了用于模式识别的 HVite 工具，其是基于 Viterbi 搜索算法的识别工具。在使用模式识别的过程中需要两个专用的配置文件：一是所需识别语音语法的网络文件；二是识别语音文件所对应的一本发音字典。因此，在这一步需要编写任务语法文件，通过调用 HParse 工具生成 HTK 可识别的底层 SLF 格式的语法网络文件，其用意即是告知辨识语音的文法结构；编写发音字典，发音字典的用意在于告诉 HTK 想要识别出来的音是哪些，并且这些音是由哪些 HMM 所组成的。训练过程所用的字典词汇量应该足够多，这样才可以训练出比较稳健的声学模型。在已有语法网络、发音字典和 HMM 集的基础上，调用 HVite 工具对测试语料进行模式识别。HTK 支持对存储型和实时输入型语音数据的识别，可以选择将识别结果保存在 MLF 文件。

2.4 模式分析

建立包含所有测试语料正确标音的参考 MLF 文件，调用 HTK 提供模型性能分析工具 HResults，将它和识别器输出的 MLF 文件进行比较，计算出识别准确率和相关参数。

3　实验结果及分析

3.1 语音库的建立

系统所用语音数据由 HTK 的 HSLab 工具进行录制，数据采样率为 16000Hz，是四个人的 400 个语音样本，其中 200 个样本作为训练集，另外 200 个样本作为测试集，包含有“0~9”十个数字，每个人对每个词分别进行十次发音，录制环境为实验室。录音时声音不能太大也不能太小，因为声音过大会影响正常发音情绪，声音过小则有用信息被忽略。

3.2 实验条件

本实验使用 3.3 版本的 HTK 进行语音识别。使用工具 HCopy 对语音数据提取 39 维的 MFCC 参数，选取音节作为识别单元，初步选定具有 5 个状态数的从左至右的 HMM 模型，斯混合分量先设置为 1，然后依次增加，直到得到满意的识别率。

3.3 实验结果

依据以上条件，遵循数据准备，模型训练，模式识别和模型分析四步，最终得到系统的性能测评结果如图 5：

```
HResults -A -D -T 1 -e ??? sil -I E:\论文实验HTK\def\ref.mlf E:\论文实验
HTK\def\labellist.txt E:\论文实验HTK\def\rec.mlf

No HTK Configuration Parameters Set

====================== HTK Results Analysis =======================
  Date: Wed Nov 10 01:22:16 2010
  Ref : E:\论文实验HTK\def\ref.mlf
  Rec : E:\论文实验HTK\def\rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=86.50 [H=173, S=27, N=200]
WORD: %Corr=86.50, Acc=86.50 [H=173, D=0, S=27, I=0, N=200]
===================================================================

No HTK Configuration Parameters Set
```

图 5　实验性能测评结果图

下面，在得到本实验结果的基础上，分别考虑选取不同识别单元（音节和声韵母），更改特征参数的维数和增加高斯混合分量的个数来与本试验结果进行分析比较。

3.3.1 两种语音识别单元的比较

改变语音识别的基本识别单元可以得到不同的识别结果，本实验分别采用音节、声韵母作为基本识别单元，在特征参数的维数（39 维）和高斯混合分量的个数（1 个）一致的情况下，进行比较实验，得到的

比较实验结果如表 2 所示：

表 2 选取两种不同语音识别单元性能测试的比较

识别单元	性能测评（%）
音节	86.5
声韵母	90.3

从表 2 中的识别结果可以得出以下结论：两种不同的识别单元(音节、声韵母)中，以声韵母为识别单元，模型的识别率比起音节模型的识别率有较大的提高。因此，相对于音节而言，可以选取声韵母作为基本的语音识别单元，以满足连续数字语音识别方向的发展，且能够很好地提高系统的识别率。

3.3.2 不同特征参数 MFCC 维数的比较

改变特征参数 MFCC 的维度可以得到不同的识别结果。本实验在选取声韵母作为基本识别单元，高斯混合分量个数取 1 个的基础上，采用不断改变 MFCC 的维度，MFCC 的维数分别取 13，26，39。检验维度的增加对识别率的影响，实验结果如表 3 所示：

表 3 改变 MFCC 维数性能测试的比较

MFCC 维数	性能测评（%）
13	82.6
26	87.8
39	90.3

从表 3 可以看出，随着特征参数 MFCC 维度的增加，识别率不断提高。当特征参数取 39 维度的 MFCC 时，得到的识别率最高。因此，实验中的特征参数取 39 维度的 MFCC 最为合适。

3.3.3 增加高斯混合分量个数的比较

增加高斯混合分量可以提高模型参数的精确度，本实验在以声韵母为基本识别单元，特征参数取 39 维的 MFCC 的基础上，逐步增加模型高斯函数的个数，检验混合分量的增加对识别率的影响。实验结果如表 4 所示：

表 4 增加高斯混合分量个数性能测试的比较

NumMixes	性能测评（%）
1	90.3
2	92.0
	94.2
5	97.7
6	95.4
7	94.6

从表 4 中可以看出：随着高斯混合分量的增加，识别率不断提高。但是提高速度随着分量的增加开始放缓，当分量增加到 5 个时，识别率不再提高，反而下降。最终，在考虑到上述所有的模型改进方案之后，使用识别率最高的 5 个高斯混合分量数目来建立模型。

从以上三个方面的比较可以看出：两种不同的识别单元（音节、声韵母）中，以声韵母作为识别单元模型的识别率比以音节模型作为识别单元获得的识别率有较大的提高。随着特征参数 MFCC 维度的增加。当特征参数取 39 维的 MFCC 时，得到的识别率最高。随着高斯混合分量的增加,识别率不断提高。但是提高速度随着分量的增加开始放缓，当分量增加到 5 个时，识别率不再提高，反而下降。最终，在考虑到上述所有的模型改进方案之后，使用声韵母作为识别单元，MFCC 维数取 39，高斯混合分量个数取 5 时即可建立识别率最高的模型。

4 结论

本文依据隐马尔可夫模型原理，利用 HTK 语音处理工具箱，实现了对数字语音识别系统的设计。同时，通过对影响系统识别率的因素：识别单元的选取，MFCC 特征参数维数的更改，高斯混合分量个数的增加，进行了比较实验，最终得到了系统识别率最高的设计组合。HTK 是一套源代码开放的工具箱，其基于 ANSIC 的模块化设计方式可以方便地嵌入到用户系统中。因此，可在 VC 环境下设计一套完整的语音识别及测试系统，并对系统实现硬件实物化，最终实现产品化，使其能真正应用于实际应用中。

参考文献

1 马峻.语音识别技术研究.哈尔滨:哈尔滨工程大学,2004.

2 Young S, Evermann G, Gales M. The HTK Book. Cambridge University Engineering Department. Version3.3, 2005.

3 http://htk.eng.cam.ac.uk

4 石现峰,张学智,张峰.基于 HTK 的语音识别系统设计.计算机技术与展,2006,(10):16-10.

5 侯周国.基于 HMM 的汉语数字语音识别系统研究.长沙:湖南师范大学,2006.

6 江官星.非特定人孤立词语音识别系统的研究.成都:西南交通大学,2006.