中文语音合成系统的设计与实现①

范会敏, 何 鑫

(西安工业大学 计算机科学与工程学院, 西安 710021)

摘 要: 为了实现机器能够发出声音, 本文设计并搭建了 HTK(HMM-Tool-Kit)平台用来实现中文语音合成系统. 采用参数合成法实现了文本到语音的合成, 并对合成系统中的文本分析、韵律控制以及语音合成的实现技术进行了详细的论述. 最后在 Linux 系统下搭建环境并进行实验, 得到了预期的结果, 实现了文本到语音的转化.

关键词: HTK; 参数合成; HTS; HMM 模型; STRIGHT 合成器

Design and Implementation of Chinese Speech Synthesis System

FAN Hui-Min, HE Xin

(School of Computer Science and Engineer, Xi'an Technological University, Xi'an 710021)

Abstract: In order to realize that machine can make a sound, this paper designs and builds the HTK (HMM-Tool-Kit) platform to realize the Chinese speech synthesis system. The parameter synthesis method realizes the synthesis from text to speech, and this paper has a detailed discussion for the implementation technology of the text analysis, prosody control and speech synthesis in synthetic system. Finally, with experiments under the environment built on the Linux, the expected results are obtained, realizing the transformation from the text to speech.

Key words: HTK; parameter synthesis; HTS; hidden markov model; STRIGHT synthesizer

随着科学技术的快速发展, 智能化设备已渗入到人类生活的方方面面, 人机交互方式也发生了一些改变, 由传统的鼠标键盘交互方式开始向智能人机交换方式转换. 语音是语言的声音, 是人类交流和传递信息最自然最直接的方式, 因此, 人机语音交互成为当前研究的热点, 特别是机器学习技术的成熟, 为语音识别、语音合成提供了技术支持.

语音合成作为语音应用技术的一部分, 被越来越多的人研究并重视. 国外对语音合成技术的研究已有几十年的历史, 近 10 多年以来"IBM", "Microsoft", "Motorola"等国际巨头纷纷投入巨大的人力、财力进行语音技术的研究, 陆续出现了英语、日语、法语、西班牙语等语种的 TTS 商品, 尤其是英语语音系统研究时间较长, 其成果已应用于多种语音翻译系统中. 国内的研究虽然起步比较晚, 但也已有数十年的研究, 如科大讯飞一直致力于研究以语音交互技术为核心的人工智能系统, 在智能语音技术领域有着长期的研究积累,并在中文语音合成、语音识别等多项技术上取得了国际瞩目的成果.

现如今对语音合成技术提出了越来越高的要求, 希望合成出来的声音能够越来越流畅、自然, 并接近原始声音. 随着统计建模方法被引入到语音合成领域, 尤其是隐马尔科夫模型的应用得到了快速的发展, 文献[2]和文献[3]中提到的使用最大似然参数生成算法来实现合成时的参数预测, 最终经过参数合成器生成语音, 但是由于传统参数合成器(如, LPC 参数合成器)的音质恢复能力较差, 这种合成方法往往不够理想[1].

针对上述合成器音质的问题, 本文在合成时采用 HTS(HMM-based Speech Synthesis System) 中 STRAIGHT 合成器的技术以提高合成的效果, 而 HTS 是严重依赖于 HTK 的. 基于 HTK 的语音合成系统主要分为两个阶段, 训练阶段和合成阶段, 训练阶段首先应该建立语音参数数据库, 这样才能通过有限的存储单元合成出无限词语, 并且能够组合出连续的语句,

① 收稿时间:2016-06-01;收到修改稿时间:2016-07-14 [doi:10.15888/j.cnki.csa.005616]

除此之外还需要创建一定的韵律规则, 这样才能对存储单元的韵律进行调整, 解决不同语言环境下的发音不同问题.

1 语音合成系统模块

语音合成系统一般主要包括三个模块: 文本分析模块、韵律处理模块以及语音合成模块, 计算机中任意的文字依次通过这三个模块就可以转换成自然流畅的语音输出.

文本分析主要是在词典的作用下对输入的文本信息的处理, 使其变为计算机能够理解的语言. 主要的工作就是将输入的文本规范化, 查找拼写错误以及过滤掉一些不发音的字符; 分析文本中词语的边界, 确定文字以及专有名词的读音.

韵律处理就是对语音进行语调、重音、时长以及停顿的处理. 对语音韵律的控制主要是通过对韵律参数, 如基频、时长、音强等的处理, 确定经过文本分析后的词语的具体发音、停顿位置及其轻重读音, 这样才有了中文语音的抑扬顿挫.

合成模块就是将经过文本分析以及韵律处理后的信息合成出声音. 目前后端的语音合成方法主要有两种: 一种是参数语音合成法, 另一种是波形拼接法. 参数合成法的基本思想是对合成单元(多采用音节、半音节或者音素)的自然语言按照一定的方法进行研究, 得到每个单元的特征参数并存储起来[2], 称之为音库; 合成时, 调用相应合成单元的特征参数并根据给出的规则变换后送入参数合成器, 得到最终的合成语音. 而波形拼接法的核心思想是直接将自然语音波形进行存储, 构成一个语音库, 合成时根据待待合成文本信息从语音库中挑选合适的波形, 再通过拼接算法将挑选出来的波形拼接在一起形成连续的自然语音.

2 基于HMM的语音合成

自 20 世纪末以来, 基于统计声学建模的语音合成技术迅猛发展, 因为它具有系统构建速度快、自动化程序高、合成效果稳定等优点, 逐渐成为语音技术领域的研究热点[3]. 其中, 隐马尔科夫模型(Hidden Markov Model)是最常用的模型, 因其系统结构简单, 基本上不需要任何语言学知识指导系统训练, 构建时间短, 构建过程基本不需要人为干涉, 因此被研究者所青睐. HMM 应用于语音信号的建模技术已经比较成熟, 基于 HMM 的语音合成技术其本质上也是一种参数合成方法.

2.1 HMM 简介

马尔科夫的定义是在已知系统目前的状态条件下, “将来”的状态与“过去”的状态无关, 只与现在的状态有关的一种模型, 可以看成是一种无记忆的单随机过程.

而假设有一个系统, 在任何时刻被认为处于有限个状态中的某一个状态下, 在均匀划分的时间间隔上, 系统的状态按一组概率发生变化, 这组概率和状态有关, 而且这个状态对应一个可观测的物理事件[4], 故称之为可观测马尔科夫过程.

隐马尔科夫过程是一种双随机过程, 该过程只能通过另一组随机过程才能观测到, 另一组随机过程产生出观测序列, 而这组的行为是不可见的. 一般使用如下五元组来描述隐马尔科夫模型:

$$\lambda=(N, M, \pi, A, B)$$

其中, N 是有限个状态数, M 是每个状态可能的观察值的数目, A 是状态转移概率矩阵(与时间无关), B 是给定状态下观察值的概率分布, π 是初始状态空间的概率分布.

2.2 基于 HMM 语音合成的系统结构

人的语言过程可以看作是一个双重随机过程, 语音信号本身是一个可观测的时变序列, 而由人的大脑中产生的语法知识和语言需要是一组不可观测的状态序列, 语音是由大脑所发出的参数流[5]. 因此, HMM 较合理的模仿了这个过程, 可以描述语音信号的整体非平稳性和局部平稳性, 是一种较为理想的语音信号模型[6]. 基于 HMM 的语音合成过程主要分为两个阶段: 训练阶段和合成阶段.

在训练阶段, 首先对存储的音库进行参数提取, 主要包括: 梅尔倒频谱系数(MFCC)和基频(F_0)参数. 然后根据提取的声学参数以及语音数据对应的标注文件为每个语音合成单元建立各自的模型, 这些模型是训练好的单音素模型, 因此还需要设计上下文属性集合的来对模型进行扩展, 再对扩展后的模型继续训练.

在合成阶段, 首先对给定的文本通过文本分析模型进行分析, 并根据文本的上下文关系在一定程度上对文本进行理解, 然后在 HMM 模型作用下进行目标模型判决并且在连续密度隐马尔科夫模型(CD-HMM)下生成参数序列, 最后将参数序列送入合成器中合成

出语音, 如图 1 所示.

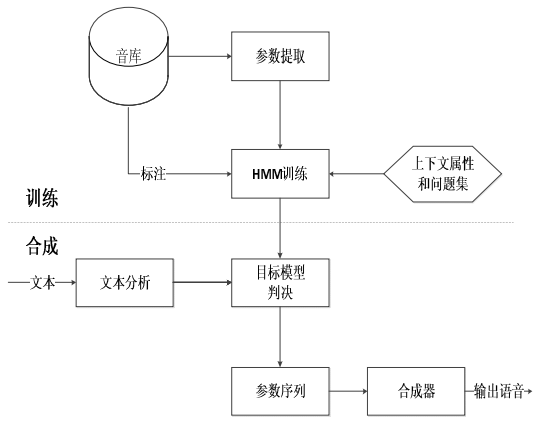

图 1 基于 HMM 的语音合成系统框图

3 基于HTK的语音合成系统

HTK(HMM Tools Kit)工具包是一个由剑桥大学开发的专门用于建立和处理 HMM 的实验工具包, 主要用于语音识别、语音合成、字符识别和 DNA 测序等应用, HTK 还是一套源代码开放的工具箱,其基于 ANSI C 的模块化设计方式可以方便地嵌入到用户系统中.

3.1 HTK 环境搭建

本文是在 Ubuntu14.04 系统下对 HTK3.4+HTS2.1 进行安装且编译, 首先必须确保系统中安装了 g++和 x11 库. HTK 安装完毕后, 运行其中自带的例子测试安装编译是否成功, 如果出现如图 2 所示的结果, 即安装编译成功.

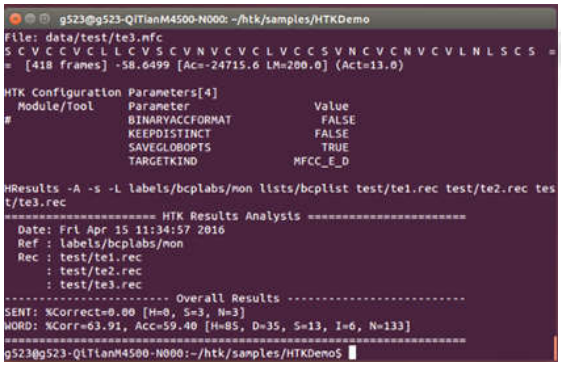

图 2 HTK 环境配置测试图

3.2 数据准备

HMM 模型的训练需要录制足够多的语音数据(称为语料库). 该录制过程可以使用 HTK 中的命令来完成, 录制的同时也要对语音进行手动标注得到文本文件, 这里的标注主要用于对语音进行时长的标注(区分静音段和语音数据段).

模型并不是使用语音数据直接进行训练, 而是根据提取的特征值进行训练, 这里使用的特征值包括频谱参数 MFCC 以及基频参数 F_0. 对于频谱多采用连续概率分布 HMM 建模[7]. 使用 HTK 的 HCopy 命令就可以完成提取 MFCC 特征的工作, 训练时特征值会作为观测值去训练 HMM 模型, 提取特征值时包括分帧、加窗、求取频谱及倒谱, 这样确保提取出的特征更加紧凑并尽可能多的保留语音的内容信息. 而对于基频 F_0 的建模, 由于清音部分是没有基频的, 即使相对于元音, 基频也是一个一维的特征, 所以, 在HTS中, 采用多空间概率分布 HMM 模型来对它建模.

3.3 模型训练

模型训练的过程其实就是一个确定 HMM 模型五元组的过程. 而合成的语音质量的好坏一定程度上取决于训练出的模型质量, 对于使用 HTS 工具训练 HMM, 关键在于一些参数的设置[8], 如定义 HMM 原型时, 其中均值和方差的取值, 以及转移概率矩阵需要给出合理的值. 通过大量的观测序列, 即语音的特征参数, 在 Baum-Welch 算法下得到模型的观察值的概率分布 B, 从而确定了该 HMM 模型.

HTK 中提供了训练工具 HRest, 使用它对所有的训练集进行嵌入式训练, 经过多次迭代优化, 最终会得到每个合成单元的 HMM 模型文件, 图 3 是对语音段中的一个词语“打开”训练得到的模型结果. 即, 经过多次迭代最终确定 HMM 模型的 5 个参数, 也就得到了训练后的模型.

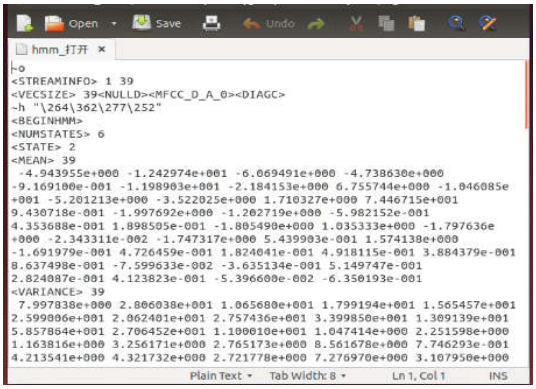

图 3 训练后的模型结果

3.4 语音合成

3.4.1 文本分析和韵律控制

对于一个待合成的文本文件, 计算机根本不能直接识别其内容, 因此需要通过一个前端的分析, 将待合成的文本信息转换为计算机所能识别的形式, 这里要想将我们所给的中文文本合成语音, 首先应该将整个文本信息分割成合适的合成单元, 在文章中选用的合成单元是声韵母, 经过该模块, 待合成的文字就会转换为拼音, 并且能够进行多音字的消歧、日期或数字的转换. 在这个过程中本文设计实现了音节的声调用以处理韵律中的声调, 以及语句的停顿位置用于处理韵律中的停顿. 如下所示是几个文本的分析示例:

文本信息“我们是中国人”, 分析后的结果为: wo3men1/shi4/ zhong1guo2/ren2

文本信息“2016/1/4”, 分析的结果是: er4/ling2/yi1/liu6/nian2/si4/yue4/yi1/ri4

文本信息“3.14”, 分析结果是: san1/dian3/yi1/si4

其中, 数字表示的是语音的声调, /表示词语的边界及停顿位置, 用来控制合成语音的韵律.

3.4.2 参数合成语音

语音合成和语音识别是一个相反的过程, 在语音识别中, 给定的是一个 HMM 模型和观测序列(也就是特征参数, 是从输入语音中提取得到), 要计算的是这些观测序列对应的最有可能的音节序列, 然后根据语法信息得到识别的文本. 而在合成系统中, 给定的是 HMM 模型和音节序列(经过文本分析得到的结果), 要计算的是这些音节序列对应的观测序列, 也就是特征参数.

通过 STRAIGHT 合成器提取的谱参数具有独特特征(维数较高), 所以它不能直接用于 HTS 系统中, 需要使用 SPTK 工具将其特征参数降维, 转换为 HTS 训练中可用的 mgc(Mel-generalized cepstral)参数, 即, 就是由 STRAIGHT 频谱计算得到 mgc 频谱参数, 最后利用原 STRAIGHT 合成器进行语音合成.

3.5 结果对比

经文本分析之后得到的文本序列在 HMM 模型下生成所需要的谱参数和基频参数, 最终进行后端的参数合成, 在 HTS 中经过 STRAIGHT 合成器最终实现将语音参数合成语音, 即得到文本所对应的语音文件. 如图 4、图 5 所示, 分别是经过 LPC 合成器和经过 STRAIGHT 合成器在 Praat 软件中对同一句话得到的语谱图.

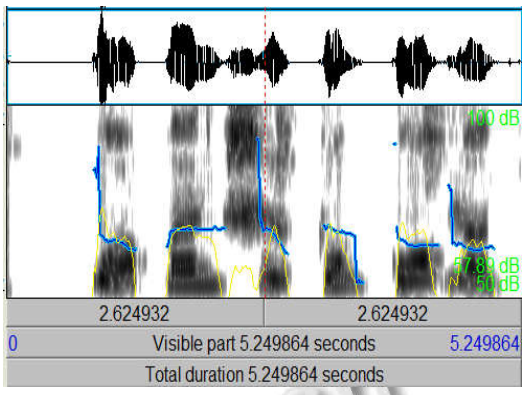

图 4 LPC 合成器合成语音的语谱图

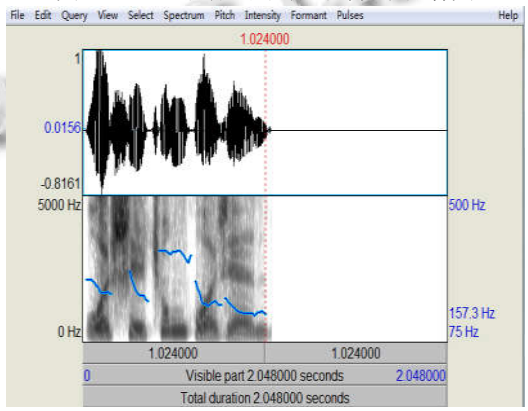

图 5 STRAIGHT 合成器合成语音的语谱图

通过比较图 4 和图 5 两幅语谱图可知, LPC 参数合成器合成的语音之间的连贯性比较差, 而使用 HTS 技术中 STRAIGHT 合成器合成的语音连续性较好, 合成的效果较理想.

4 结语

语音合成技术是完成人机语音通讯不可或缺的重要部分, 随着人工智能技术的快速发展, 语音合成的应用也越来越广泛. 文章在分析了 HMM 的原理后, 根据其在语音合成方面的应用进行了详细的原理分析, 并且搭建了 HTK 环境, 对语料库的参数进行训练, 利用 HTS 技术在 STRAIGHT 合成器中最终合成了语音, 实现了从文本到语音的转换, 相比较使用传统的合成器合成的语音连续性较好, 合成的效果较理想. 近年来, 基于波形拼接的语音合成得到了研究者们的大量关注, 因其使用原始的自然语音作为语音库, 使得合成的结果更加接近人的原始声音[9], 效果更加的流畅, 所以在后期的工作中会对此方法进行深入的研究, 希望合成的语音效果越来越自然.

参考文献

1 王仁华,戴礼荣,凌震华,胡郁.基于统计建模的可训练单元挑选语音合成方法.科学通报,2009,(8):1133–1138.

2 刘浩杰,杜利民.语音合成技术的发展与展望.微计算机应用,2007,28(7):726–730.

3 宋阳.基于统计声学建模的单元挑选语音合成方法研究[硕士学位论文].北京:中国科学技术大学,2014.

4 李婧.基于 UBM 的发音质量评价系统的设计与实现[硕士学位论文].天津:南开大学,2007.

5 顾香.面向统计参数语音合成的方言文本分析的研究[硕士学位论文].兰州:西北师范大学,2014.

6 王婧,朱黎.一种基于改进的 LPC 参数倒谱分析的说话人识别方法.大众科技,2008,(8):28–29.

7 赵建东,高光来,飞龙.基于 HMM 的蒙古语语音合成技术研究.计算机科学,2014,41(1):80–82.

8 纪正飚,王吉林,赵力.基于 HMM 的中英文语音合成技术研究.科学技术与工程,2014,14(32):237–240.

9 雷鸣.统计参数语音合成中的声学模型建模方法研究[博士学位论文].合肥:中国科技大学,2012.