基于混合基元模型的非定长基元选取算法

倪 昕, 蔡莲红

(北京 清华大学计算机系, 北京 100084)

E-mail: nx01@mails.tsinghua.edu.cn; clh-dcs@tsinghua.edu.cn

摘 要: 介绍了面向中英文双语应用的英文语音合成系统中基于混合基元模型的非定长基元选取算法. 清华大学计算机系人机语音交互实验室针对中英文混读相同发音人的限定, 实现了基于混合基元模型的语料库构建和鲁棒灵活的非定长基元选取方法, 在一定程度上弥补了发音人英语发音不饱满、自由度大的缺陷, 真正实现了相同发音人中英文混读的要求. 试验证明, 采用这些方法能够极大的提高合成质量, 达到令人满意的效果.

关 键 词: 文语转换(TTS); 混合基元模型; 非定长基元选取; 语料库

中图分类号: TP391 **文献标识码**: A **文 章 编 号**: 1000-1220(2005)06-1079-04

Hybrid Unit Model Based Non-uniform Unit Selection

NI Xin, CAI Lian-hong

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: This paper presents hybrid unit model based non-uniform unit selection in Chinese and English bilingual application oriented English text-to-speech system. In order to meet the speaker-consistent requirement and address the nonnative speaker problem which manifests deficiency and unrestraint in pronunciation, the Human Computer Speech Interaction Laboratory at Tsinghua University constructs acoustic inventory based upon hybrid unit model and implements an efficient, robust and flexible non-uniform unit selection method. All these prove effective experimentally in dealing with the nonnative speaker problem and in promoting the overall synthesis quality.

Key words: text-to-speech (TTS); hybrid unit model; non-uniform unit selection; acoustic unit inventory

1 引 言

当前, 英语作为一种国际语言, 已经深入到我们的生活中, 这给语音合成的研究带来了新的挑战. 客观上要求文语转换(TTS)系统能够实现相同发音人的英汉双语, 甚至多语种的合成和混读. 近年来汉语语音合成技术得到了长足的发展, 已经进入实用阶段; 与汉语相比, 英文合成的研究起步较早, 水平较高. 然而在中英文混读的背景下, 中、英文系统结合却存在一定困难: (1) 国内对英语 TTS 系统的研究尚处于起步阶段, 水平有限. (2) 相同发音人的要求难以满足. 具体来讲, 合成系统的最终效果和录音人的发音水平密切相关, 因此对发音人的要求比较苛刻[2], 需要经过专业训练的广播员. 研究中英双语合成时, 这一限定所带来的问题尤为突出. 现有的中文合成系统的发音人大都非英语母语, 即使经过必要的训练, 发音不饱满、不规范的现象也难以避免. 在保持原有中文系统不变的前提下, 非母语发音人问题已经成为双语甚至多语种合成研究必须解决的关键问题.

国内外研究机构在此方面已做了很多尝试. 国内的一些研究单位限定了合成内容, 仅合成预先录制常用英语单词; 另外某些单位基于原有中文合成系统以音节为基元合成任意英文文本, 但合成效果不理想, 音节停顿或跳跃明显, 缺乏必要的连读发音和流利度; 国外的某些商业系统则舍弃了相同发音人的要求, 集成了音色相近的不同发音人的中、英文音库实现双语混读, 然而音色差异仍存在并会导致双语合成时音色的跳变. 事实上解决双语合成发音人问题的根本途径是使发音人经过严格、专业、系统的学习和训练, 中、英文都达到专业发音水平. 然而考虑到多语种合成的需求, 使发音人同时具备多种语言专业发音水平是不实际也是不可能的. 因此, 必须从语音合成技术本身出发, 寻求发音人问题的解决方法. 美国 Bell 实验室此前曾在多语种语音合成方面做了有益的尝试[8].

由于语言限制和侧重点的不同, 目前国内对英文语音合成还缺乏细致深入的研究. 清华大学计算机系人机语音交互实验室以英国爱丁堡大学 Festival[1]系统为框架, 建立了一套面向中英双语混读的英文系统试验平台. 本文是在此背景下对英文合成做出的一点粗浅尝试. 特别是针对非母语发音人发音问题提出了一些新的技术手段, 包括语料设计、标注处理以及非定长基元选取模型等, 较为有效地减少了发音人问题给合成系统带来的影响.

2 混合基元模型

基元模型确定了拼接合成的基本单元, 大到词, 小到半音素(half phone), 都可以作为拼接单元. 考虑到连续语流中的韵律变化以及协同发音的影响, 基元模型在很大程度上影响合成语音的质量. 一般来讲, 基元大, 效果好, 但基元数量也越多, 音库庞大且难以覆盖所有基元; 基元小, 灵活性好(比如半

收稿日期: 2003-12-28 基金项目: 国家"八六三"高技术项目(2002AA1170)资助; 国家自然科学基金(60275014)资助 作者简介: 倪 昕, 硕士研究生, 研究方向为基于语料库的英文语音合成; 蔡莲红, 教授, 博士生导师, 研究兴趣在语音合成, 多模态生物特征识别等.

音素可以根据需要拼接成音素或者 diphone[3]), 但拼接点增多可能导致质量损失 表 1 比较了当今英文合成系统普遍使用的几种基元模型

表 1 基元模型

基元模型	代表系统	数量	备注
音素 phone	CHATR [4]	少于 50	协同发音强
双音素 diphone	Bell lab TTS	1500- 2000	协同发音较弱
半音素 half phone	AT&T NextGen	少于 100	灵活性高
三音素 triphone	BT Laureate	多	协同发音弱
音节 syllable	一些国内系统	超过 10000	协同发音弱
半音节 demi-syllable	Telcordia Orator	800 initial, 1200 final	协同发音弱
单词 word		过多	适于限定域合成

实际上, 基元模型的确定是在同时考虑了音库规模和涵盖诸如协同发音等各种语言现象这两个问题之后做出的折衷选择 如上所述, 长基元可以涵盖各种音素组合产生的协同发音现象, 从而减弱其影响, 但同时意味着基元数目的增多和语料库规模的增大 对于英文来讲, 其发音中弱化、浊化、连读、失暴等语言现象非常普遍, 解决这些问题的方法通常是由语言学家总结规律, 在音库中收入多种音位变体, 涵盖各种可能的音变现象, 使得合成结果更为自然流畅 Bell 实验室多语种 TTS 系统[5, 8]就以双音素(diphone)基元为主, 对不能涵盖的协同发音现象(如元音- 爆破音- 元音组合), 还收集了上下文敏感的音位变体(allophonic)、三音素(triphone)、甚至是单词(常用的功能词). 相比之下, Festival 仅以双音素为基元, 缺乏不同基元的相互补充, 因此合成自然度有限 然而, 这种办法存在一定局限: 连续语流中语言现象复杂多变, 并非完全依规律而行; 非母语发音人发音自由度大, 若无严格训练则情形更为严重; 不同发音人的发音习惯和规律都有所不同, 因此人工规则的总结不但需要丰富的语言学知识而且还与发音人相关; 人为加入多种模型将破坏音库的一致性, 增加工作量和复杂度

本文在已有的基元模型基础上提出了混合基元模型的概念 其基本思路是以语料库为基础, 将各种基元模型与非定长基元选取结合在一起, 在基元选取过程中依据全局代价动态确定、生成拼接基元 混合基元模型涵盖了从半音素到单词、短语的所有可能, 使基元模型的运用更具系统性和灵活性

依据这一思路, 本文首先根据基元边界的不同将现有的基元模型进行了归类 为说明方便, 我们把以音素中部稳定段作为拼接点的基元拼接称为稳定型拼接, 把以音素边界过渡段作为拼接点的称为过渡型拼接 据此可把基元分为三类: 过渡边界型、稳定边界型和混合型 过渡型的基元边界均属于过渡型拼接, 比如表 1 中的音素基元和音节基元; 稳定型的基元边界都属于稳定型拼接, 如双音素基元; 混合型基元的两个边界则分别是稳定型和过渡型, 比如半音素和半音节基元 已有的研究表明, 一般情况下稳定型拼接的平滑度要优于过渡型拼接[6]. 然而, 稳定型基元往往要考虑前后音联组合, 因此基元数量众多, 每个基元的样本量少, 表现出数据稀疏性, 难以满足韵律变化的选音要求 以典型的音素(过渡型基元)和双音素(稳定型基元)为例, 两者基元长度大体相同, 在音库规模相同的情况下, 双音素基元由于数量众多, 每种基元的样本数量要比音素显著稀少, 甚至会出现一定数量的缺失; 音素基元虽然属于过渡型拼接, 受前后音素协同发音影响较大, 但待选样本数量多, 通过选音算法的控制亦可弥补拼接的不足 因此, 如何最大程度利用稳定型拼接, 而又能弥补数据稀疏性问题是本文提出混合基元模型的一个出发点 同时非母语发音人发音不饱满、自由度大, 若用统一的基元模型和拼接规则来刻画, 那么同一基元的不同样本的频谱则表现出非常显著的差异, 对合成质量影响很大 综上所述, 以稳定型拼接为主, 过渡型拼接作为前者的补充和退化形式, 通过选音算法综合运用各种基元模型是本文解决非母语发音人问题的基本思路

进而, 不同类型基元的生成通过分割和归并基本音素单元来实现 在构建音库时, 系统对每个音素单元都标注了前后过渡点和稳定点 在每个标注点上都可以进行基元的切分和归并, 例如前后相连的两个音素分别从稳定点切开、在相邻的过渡点归并就构成了一个双音素基元 这种音素单元的切分和归并是在基元选取过程中依据全局代价动态进行的, 无需人工规则的指导 具体过程将在基元选取部分讨论

3 非定长基元选取

基元选取问题可以归结为三点: (1)影响语音韵律特征的因素有哪些, 特征参数如何构造; (2)如何衡量待选基元与目标基元的差距, 代价函数如何构造; (3)如何实现快速有效的算法而不明显降低选音质量

针对以上问题本文构造了上下文索引树来快速寻找待选音素单元, 通过 Viterbi 搜索计算全局匹配代价, 对音素单元进行归并或切分, 动态生成合成基元

3 1 上下文索引树

在大规模语料库合成系统中, 树形结构常被用来细分基元类别, 减少候选基元数量, 实现快速索引和的目的 [7]中使用决策树(CART)方法对具有相似韵律特性的基元进行聚类, 这种聚类方法上下文特征不突出, 难以实现相连单元归并、相似单元切分的功能, 因此不适于混合基元模型 本文对基于决策树的选音模型进行了试验, 结果表明采用这种方法的基元连续性不超过 10%, 拼接点较多而影响合成质量

面向混合基元模型的上下文索引树方法能够有效的解决这一问题 使用索引树的主要目的是对基元库进行预选取, 削减候选基元样本数量, 提高合成效率 索引树结构见图 2

上下文索引树依据上下文特征对音素单元进行分级索引 目前划分了 3 个级别: 0 级代表独立音素集; 1 级为 0 级的子集, 集聚了具有相同双音素类的音素集; 2 级为 1 级的子集, 集聚了具有相同三音素类的音素集 索引项即为音库中音素单元 索引树的级别具有自上而下的包含关系, 并可根据语料库规模调整层级结构, 优化基元选取效率 特别是, 以上下文特征进行语音单元索引, 特别有利于最终合成基元的生成与归并 在选音时, 依据上下文特征优先选用级别最高的单元

集, 若高级别单元集为空或样本数量过少则向上逐级遍历索引树, 直到目标单元集满足要求为止. 索引树的构造也同时体

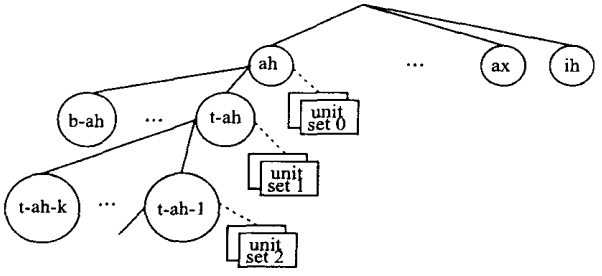

图 1　上下文索引树

现了选音过程所要求的鲁棒性和灵活性.

3.2　音素特征向量与代价度量

本文用音素特征向量表达音素单元的韵律和语境特征. 目前使用了 14 维特征, 包括:

- 音联信息: 前音素、后音素
- 重音信息: 音节重音(stress)、音节重读(accent)
- 位置信息: 音节中的相对位置(onset/coda)、前音素的音节相对位置、后音素的音节相对位置、音节中的绝对位置、是否音节首、是否音节尾
- 停顿信息: 后停顿类别、前停顿类别
- 韵律信息: 时长、平均基频

其中重读(accented)是区别于词法重音(stress)的逻辑重音或语调重音; 音节中的相对位置有两个值, onset 表示该音素处于所在音节中的元音前, coda 表示音节元音或处于元音后; 停顿类别分为四类, 无停顿 1(词中音节)、无停顿 2(非短语末词尾音节)、停顿(短语末词尾音节)、长停顿(长短语末或句末词尾音节).

本文选音代价函数的定义与[4]相同, 以音素的特征向量和对应权重向量计算匹配代价, 以拼接类型和平均基频计算拼接代价, 匹配代价和拼接代价经过归一化和加权求和最终获得全局代价. 其中权重是在设置初始值后通过人工指导下的机器学习方法获得[9]: 即对现有的选取结果进行人工打分, 然后根据评测结果自动修正权重. 通过权重的调整可以进一步弥补发音人的不足, 去粗取精.

3.3　基元生成与归并

采用上下文索引树获得候选单元并计算匹配代价后, 基元的生成与归并通过 Viterbi 算法搜索最佳路径实现. 如前所述, 基元拼接可分为稳定型和过渡型, 因此, 前后音素(Ui, Uj)拼接时有以下三种情况:

1. 理想拼接: Ui 和 Uj 在音库中相连. 两音素归并, 拼接代价为 0.

2. 稳定型拼接: 音库中与 Uj 相邻的前一单元 Uk(如果存在)与 Ui 相同(例如同为 ah). 拆分 Ui 取前半段, Uk 拆分取后半段并与 Uj 合并, 拼接代价加权系数取小. 同时更新 Ui 的目标代价, 使其成为 Ui 与 Uk 目标代价的综合.

3. 过渡型拼接: 音库中与 Uj 相邻的前一单元 Uk 不存在或与 Ui 不同. 拼接加权代价系数取大.

图 3 显示了基元生成与归并的一个例子: 合成单词“return”. 框内为生成的基元(归并的结果), 框间为基元拼接. 第一行为稳定型基元拼接, 第二行为过渡型基元拼接. 同时我们可以看到合成方案中产生了不同的基元类型, 这使得选音算法更加灵活, 适应于发音人自由度较大的情形.

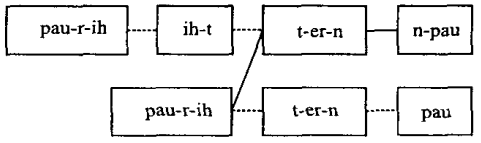

图 2　基元生成与归并

实际上, 在 Viterbi 搜索过程中, 基元的生成和归并会存在多种方案. 如图 3 显示了合成“return”的两种极端方案, 第

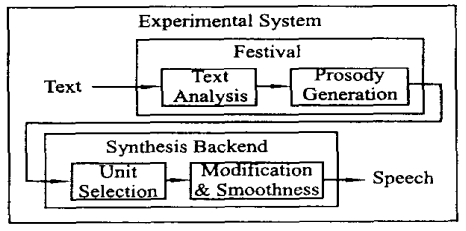

图 3　系统结构

一行基元均为稳定边界型, 拼接代价小, 而拼接点多于第二行的过渡边界型. 最终的方案则由全局代价最小的最佳路径确定, 可能同时混有多种基元类型, 如图中实线所连方案. 通过对基本音素单元的搜索和拼接基元的动态拆分和归并, 既可以最大限度地使用稳定型拼接提高合成结果平滑度, 又可以充分利用音库数据, 降低数据稀疏性的影响; 同时对音库规模也具有较好的适应性, 如移动设备小型音库覆盖的双音素、三音素组合有限, 选音过程会自动退化成对音素单元的选取. 最后需要指出的是, 虽然系统的语料设计仅仅考虑了双音素、三音素基元的收录, 但收录文本中常用音节、单词的出现频率较高, 有利于基元选取中长基元(比如音节、单词)的生成, 提高合成质量.

4　测试系统

4.1　语料设计

语料设计应面向韵律变化, 用尽量少的语料, 覆盖尽可能多的语言现象. 试验系统的语料面向混合基元和非母语发音人, 具有以下特点:

1. 从真实文本中抽取语料, 依优先顺序覆盖双音素基本集、双音素扩展集、三音素集.

2. 文本选自英语教程, 优先收录常用高频词.

3. 为便于录音, 以短语为单位(通常 5- 10 词)收录语料.

其中双音素基本集是基于美语的 DARPAbet 音素集[1]之上的双音素集, 包含 cvc, cv, vc, vv, cc, cc-cluster, -sil, sil-等各种元音、辅音、辅音簇和静音组合. 扩展集在此基础上考虑了音节重音(stress), 如(t-aa 0 1)表示 t 所属音节不重读, 而 aa 所属音节重读. 扩展集的引入, 是考虑到重音属性在表述基元韵律时格外重要, 而其声学表征又不完全明确.

系统语料设计部分以 Festival 的前端文本分析模块为基础, 将长句切割成短语, 通过文本分析对短语中各种基元的语境、韵律信息做出标识, 避免具有相似特征的基元被重复收录. 通过对几种基元集合的综合考虑, 试验系统测试语料共收集短语约 2000 条, 其中双音素基本集覆盖约 85%, 扩展集接近 60%, 三音素集亦有一定覆盖. 初步构建起面向混合基元模型的语料库.

4.2 音库构建

试验系统音库共录制短语语料 1000 条. 为引导发音人正确读音, 提高录音效率, 语料事先用 Festival 合成了提示音. 录音要求所有音素发音与提示音保持一致, 注意单词连读, 韵律尽量自然平缓, 语速稍慢. 基元切分和基频标注均以 Festvox [2]为基础自动处理. 其后对切分结果进行了人工检查和修正. 由于合成基元是在选音过程中动态产生, 因此建立音库时省去了将所有基元类型一一标出的工作, 只需标出每个音素的起始点(时间)、稳定点和结束点, 音库格式统一、简单. 下面是音库中的某一音素索引项:

ah_0 ph_1009 1.726980 1.784250 1.831240 p b 1 coda coda onset 0 1 0 0 0 1 0. 10426 214. 3

其中 ah_0 是音库中的音素标识, ph_1009 代表录音文件名, 后三项分别对应该音素的起始点, 稳定点和结束点, 最后是音素特征向量. 此外, 针对发音人问题还进行了一些特殊标注, 如发音是否饱满, 音位关联度等. 这些标注有利于选音模块针对发音人问题做出正确决策. 试验证明这起到了良好效果. 除个别标注外, 音库构建都可自动进行, 规模也可随意调整.

4.3 系统结构

本文以 Festival 为框架, 引入新的后端合成模块(包括音库构建、非定长基元选取以及韵律修改与拼接平滑三部分), 建立了一个基于大规模语料库的拼接合成系统. 系统结构如图 3 所示.

5 评 测

试验音库共录制短语近 1000 条, 总长约 45 分钟, 16KHz 采样, 约 100M 数据. 通过主观评测, 其效果优于一般的基于双音素的拼接合成系统(Festival)和基于决策树的聚类选音系统[7](与本系统使用相同的录音音库). 图 4 显示了十位评测人对任选语料合成评分的平均结果.

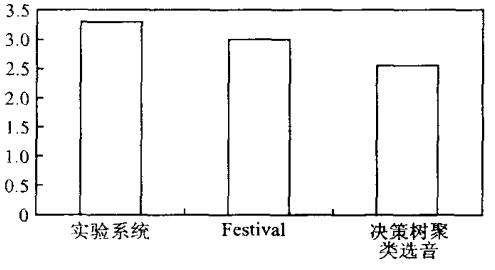

图 4 主观评测

此外, 我们还对上下文索引树的性能做了简要评测. 使用上下文索引树可以大幅度削减候选基元数量, 十倍量级地提高系统性能, 且不明显降低合成质量. 我们对任意选取的 1000句文本合成语音的结果显示, 使用和不使用上下文索引树所生成基元的一致性达到 87.2%, 而在不一致的情况中, 使用树所生成基元的长度往往大于不使用的情形, 拼接点个数也更少. 图 5 显示了使用索引树合成 1000 句文本的基元长度的统计结果.

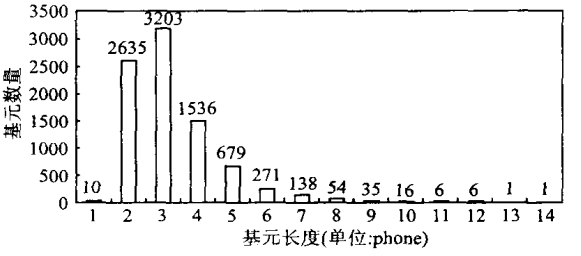

图 5 基元长度统计

6 结 论

本文介绍了清华大学计算机系人机语音交互实验室在面向中英文混读的英文 TTS 系统研究中所采用的非定长基元选取算法. 解决非母语发音人的问题, 提高发音人发音水平仅是一个方面, 本文从试图从另一方面, 即从系统角度寻求解决问题的方法. 试验表明, 采用基于混合基元模型构建语料库和鲁棒灵活的非定长基元选取方法, 在一定程度上能够弥补发音人英语发音不饱满、自由度大的缺陷, 真正实现中英文混读相同发音人的要求.

英语在语言、语音学上都有很多独特之处. 缺乏英语语音学知识指导, 非母语研究人员以及非母语发音人等因素都是国内英语语音合成研究者所要面临的困难. 本文是对英文合成研究做出的一点粗浅尝试.

References:

[1] Black A, Taylor P. The festival speech synthesis system: System documentation [EB/OL]. http://www.festvox.org, 2003.
[2] Black, Alan W, Lenzo Kevin A. Building voices in the festival speech synthesis system [EB/OL]. http://www.festvox.org, 2003.
[3] Alistair Conkie, Robust unit selection system for speech synthesis [C]. In Procceedings of the 137th meeting of the Acoustical Society of America, 1999.
[4] Andrew J. Hunt, Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database [C]. In: Procceedings of International Conference on Acoustic, Speech, and Signal Processing, 1996, 373-376.
[5] Sproat R, Olive J. A modular architecture for multi-lingual text-to-speech [C]. In: Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, 1994, 187-190.
[6] Beutnagel M, Conkie A, Syrdal A. Diphone synthesis using unit selection [C]. In: Procceedings of the 3rd Escacocosda International Workshop on Speech Synthesis, 1998, 185-190.
[7] Black A W, Taylor P. Automatically clustering similar units for unit selection in speech synthesis [C]. In: Eurospeech97, 1997, volume 2, 601-604.
[8] Richard Sproat, Multilingual Text-to-Speech Synthesis: the bell labs approach [M]. Kluwer Academic Publishers 1998.
[9] Wu Zhi-yong, Cai Lian-hong, Tao Jian-hua. Speech unit selection based on mandarin prosody parameter [C]. In: Proceedings of National Conference on Man Machine Speech Communication 6, 2001, 199-202.

附中文参考文献:

[9] 吴志勇, 蔡莲红, 陶建华. 基于汉语韵律参数的语音基元选取. 第六届全国人机语音通讯学术会议论文集[C]. 2001, 199-202.