软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(Suppl.(2)):63−69

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

语音合成中基于稳定段边界的不定长基元选取*

王 欣[1,2], 吴志勇[1,2], 蔡莲红[1,2]

[1](清华大学-香港中文大学媒体科学、技术与系统联合研究中心(清华大学 深圳研究生院),广东 深圳 518055)

[2](清华信息科学与技术国家实验室(清华大学),北京 100084)

通讯作者: 吴志勇, E-mail: zywu@sz.tsinghua.edu.cn

摘 要: 语音合成技术是人机言语交互中重要的媒介方式,基元选取算法一直是拼接式语音合成中的研究重点.在传统的语音合成中基于代价函数的拼接合成基元选取算法的基础上,将双音子(diphone)的稳定段边界模型应用到单词和音节中,最后使用 3 种基元模型的分层不定长选音算法,从语料库中优选出最佳合成基元序列拼接合成最终语音.该算法一方面利用分层统一的不定长选音策略,尽可能地选取具有更好韵律特性和声学连续性的较大基元,从而显著减少拼接点,将有可能发生协同发音或者切分错误的拼接点包含到更大的基元内部;另一方面通过稳定段切分修改传统拼接基元边界类型,充分利用了 diphone 的稳定段边界良好的拼接特性,从而提高了合成语音的连续性和自然度.评测结果显示,这种方法与传统 diphone 拼接合成方法相比,其合成效果有显著的提升.

关键词: 语音合成;TTS;diphone;稳定段边界类型;分层不定长基元选取

中文引用格式: 王欣,吴志勇,蔡莲红.语音合成中基于稳定段边界的不定长基元选取.软件学报,2014,25(Suppl.(2)):63−69. http://www.jos.org.cn/1000-9825/14024.htm

英文引用格式: Wang X, Wu ZY, Cai LH. Stable boundary-based non-uniform unit selection in speech synthesis. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(2)):63−69 (in Chinese). http://www.jos.org.cn/1000-9825/14024.htm

Stable Boundary-Based Non-Uniform Unit Selection in Speech Synthesis

WANG Xin[1,2], WU Zhi-Yong[1,2], CAI Lian-Hong[1,2]

[1](Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems (Graduate School at Shenzhen, Tsinghua University), Shenzhen 518055, China)

[2](Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing 100084, China)

Corresponding author: WU Zhi-Yong, E-mail: zywu@sz.tsinghua.edu.cn

Abstract: Speech synthesis technology plays an important role in human computer interaction. Based on the traditional cost function based unit selection method, this paper proposes an approach that incorporates diphone's stable boundary model into word and syllable, and utilizes multi-layer Viterbi algorithm for selecting the best path from the corpus to generate the final waveforms. With the proposed multi-layer non-uniform unit selection algorithm, the new method can not only choose the longer prosody units which have correct acoustical characteristic to reduce the concatenate points while including the potential coarticulation and bad labeled phones inside the longer units, but also fix the traditional unit boundary type to absorb the diphone's good stable joint character to improve the continuity and naturalness at concatenate boundaries. The evaluation results show that by using this approach, the synthetic speech can achieve great improvements on both naturalness and intelligibility compared with the traditional diphone-based unit selection approach.

Key words: speech synthesis; TTS; diphone; stable boundary type; multi-layer non-uniform unit selection

语音合成是使计算机能够像人一样自然说话的一种技术.目前,语音合成技术相关的研究主要集中在文语转换系统(text-to-speech,简称 TTS)这一阶段.几十年来,随着语音技术和相关理论的不断发展,合成语音的质量已经远远超越了人们最初对于音质可懂度和清晰度的需求,在自然度上也有了长足的进步.语音合成技术已经

* 基金项目: 国家自然科学基金(60805008, 61003094, 61375027, 61370023)
收稿时间: 2013-06-15; 定稿时间: 2013-08-21

趋于成熟和实用,开始从实验室走向人们的日常生活.近几年来,特别是随着个人智能助手 siri 在 iphone 手机上取得的巨大成功以来,人们逐渐意识到语音技术,特别是语音合成技术在人机交互中起着重要的媒介作用.语音合成技术具有广阔的应用前景.

然而,与真正的人类自然语音相比,目前的合成语音在自然度和流畅度上,仍然有所欠缺.如何进一步提高合成语音的音质、从而进一步提升人机交互体验成为当前语音合成技术的焦点问题.

一个标准的 TTS 包括文本分析前端和合成后端.目前的主要合成方法分为基于大规模语料库的基元选取拼接合成方法[1−3]和基于统计模型的参数合成方法[4,5].其中拼接合成方法研究起步较早,如今发展和应用的也更为成熟和广泛.这种方法利用一个预先录制的适当规模的人类自然语音构成的语料库,其中包含大量具有不同声学和韵律特征的语音基元,然后通过基元选取算法从中挑选出最符合人类发音特征的语音片段序列并将之拼接为一个整体,最后通过少量的语音信号处理方法进行适度的平滑和修改后合成最后的语音.这种合成方法直接利用语料库中的原始语音进行拼接合成,最大限度保留了语音的原始特性,使得合成语音的自然度大幅度提高.但是由于自然语流丰富的动态性和不同发音单元之间协同发音的影响,如果所选基元不符合发音环境或者拼接点过多导致语音连续性下降则会显著降低合成语音的音质.于是如何充分利用语料库、尽量减少对原始语音的人工修改、选出最符合合成要求的语音基元进行拼接合成,对语音合成的质量具有决定性的影响.另外一种参数合成方法则利用隐马尔可夫-高斯混合模型(hidden Markov model-gaussian mixture model,简称 HMM-GMM[6])或者隐马尔可夫模型-深度神经网络(HMM-deep belief network,简称 HMM-DBN[7])等机器学习方法对大规模人类自然语音进行统计建模,然后利用学习模型生成声学参数合成语音.这种建模方法较传统拼接合成方法在合成语音的连续性和灵活性上有较大突破,但是在音质上,仍然和自然语音有所差距,有待进一步发展.

在基于拼接的英文 TTS 中,主流的选音方法有基于分类回归树(classification and regression tree,简称 CART)的决策树聚类方法[8]和基于代价函数的选音方法[1].而主要的拼接基元则根据英文的发音特点和不同的应用场景需求,分为半音素(half-phone)、音素(phoneme)、双音子(diphone)、半音节(demi-syllable)、三音子(triphone)、音节(syllable)[9]或者更大的基元比如单词(word)甚至不同大小基元的混合基元模型[10,11]等.在这些拼接基元中,diphone 因为其良好的拼接特性和适中的基元规模得到了最为广泛的应用.著名的多语种语音合成系统 Festival[12]就是以 diphone 作为其唯一的拼接单元,并且取得了良好的效果.但是同时我们可以看到,由于 diphone 的基元过小(跟 phoneme 相当),会导致最终合成语音中拼接点过多(基本等同于一句话中的音素数目),而过多的拼接点则会影响最终合成语音的连续性,甚至会降低合成语音的自然度.

本文在代价函数的基础上,提出了一种新的基元选取算法.该算法一方面利用分层统一的不定长选音策略,尽可能地选取具有更好韵律特性和声学连续性的较大基元,从而显著减少拼接点;另一方面通过稳定段切分修改传统拼接基元边界类型,充分利用了 diphone 良好的拼接特性,从而提高合成语音的连续性和自然度.评测结果显示,这种方法与传统 diphone 拼接合成方法相比,其合成效果具显著的提升.本文第 1 节对拼接基元选取和边界修改方法进行阐述.第 2 节详细描述所提出的新基元选取算法中的不定长选音策略.第 3 节对该方法和实验结果进行人工评测.第 4 节进行总结和讨论.

1 拼接基元修改

1.1 协同发音

在拼接合成中,协同发音现象是指发音时声道中的不同发声部位形成阻塞,导致前位音段易于向后接音段的过渡而发生发音变化的语音学现象.这类现象主要发生在音素间的辅音-元音、元音-元音、和不发音的 r-元音连缀上[13].协同发音引起的音连会导致两个相邻音的交织重叠,不利于基元的切分和拼接,但同时它又是语音中极富动态性和表现性的特性.

连续语流中的协同发音现象只能控制而不能消除,减小协同发音影响的关键点是发掘潜在的拼接点,避免不正确和非必须的切分导致强协同发音对频谱连续性和拼接自然度的影响.表 1 比较了当今英文合成系统中

常用的几种基元类别.

Table 1 General unit models[13]

表 1 常见基元模型[13]

基元模型	代表系统	数量	备注
音素 phone	CHART	少于 50	协同发音强
双音子 diphone	Bell-Lab TTS, Festival	1 500~2 000	协同发音较弱
半音素 half-phone	AT&T NextGen	少于 100	灵活性高
三音子 triphone	BT laureate	简单组合超过 60 000	协同发音弱
音节 syllable	一些国内系统	超过 10 000	协同发音弱
半音节 demi-syllable	Telcordia Orator	800initial, 1 200final	协同发音弱
单词 word	航班报时系统,天气预报系统	过多	仅用于限定领域

1.2 diphone的声学特征

图 1 是常用于英语拼接合成系统中单词以下层级基元的定义.

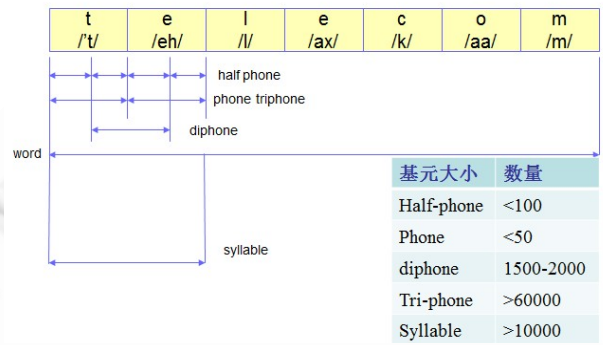

Fig.1 Definition of concatenative units

图 1 拼接基元定义

目前大多数的 TTS 系统都采用双音子(diphone)作为合成基元.这与双音子的结构有关.由图 1 可知,双音子跨越了相邻两个音素的发音区间,以第 1 个音素的发音稳定段作为开始,包含了相邻音素的发音过渡段,以后一个音素的稳定段结束.双音子模型认为协同发音现象主要出现在音素边界上,即相邻单音素的拼接点上,而在音素中部稳定段的发音比较稳定,所以双音子选择协同发音较为弱的音素稳定段作为拼接点.

1.3 拼接基元选取和基元边界修改

综合考虑英文发音特征以及基元规模,本文最终选取了双音子、音节以及单词作为基本的语音合成基元.对英文中出现较为频繁的功能单词以及常用音节(前后词缀、固定发音搭配形式等),尽量采用大基元来减少拼接点和增加连续性,降低协同发音的影响.而对于不同基元之间的过渡部分或者语料库中不存在的基元,则根据英文发音特点将其拆分为连续的 diphone 序列.

传统的语料库构建技术采用基于单音的语音识别强制对齐算法,在单音的过渡段(即前一个音素的结束和后一个音素的开始点)切分,将连续语流切分为一系列包含时间标注的单独音素序列.通过单音的标注,可以很方便地得到其上层发音基元的边界信息.但是这种边界信息都是过渡段切分,与 diphone 的稳定段切分不一致,且过渡段拼接受切分错误或者协同发音的影响较为严重,频谱和能量、基频等声学参数变化较为剧烈,容易导致拼接的不连续,所以不宜直接用来拼接.

在本文中,受到 diphone 的稳定段切分启发,我们在单词、音节的首尾音素没有使用传统的过渡段切分,而是在边界音素的稳定段切分.然后在实际拼接算法中,动态地在这些基元前后插入相应的 diphone 来保证前后基元之间的稳定段拼接.这种方法一方面能够充分利用大基元、将其与 diphone 的不同边界类型统一起来,使得不定

长选音算法可以更为简洁地实现;另一方面,类似于 diphone 的稳定段拼接将会减小拼接点的失真现象.一个稳定段基元切分如图 2 所示.

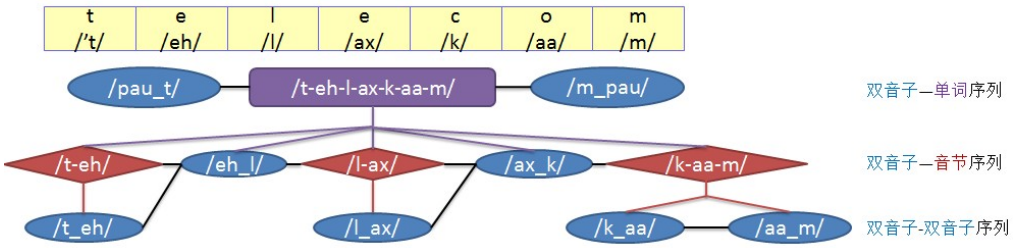

Fig.2 Stable boundary unit segmentation

图 2 稳定段基元切分

2 基于稳定段边界的不定长基元选取

2.1 代价函数

代价函数 $c(u_1,u_2,...,u_n,t_1,t_2,...,t_n)$用来评价从语料库中选出的候选基元是否符合合成要求以及所选序列是否满足语音学上的连续性,包括匹配代价和拼接代价两部分,见式(1):

$$C(u_1,u_2,...,u_n;t_1,t_2,...,t_n)=\lambda\times\sum_{i=1}^{n}\frac{\sum_{j=1}^{k}w_j^{t\arg et}c_j^{t\arg et}(u_i,t_i)}{\sum_{j=1}^{k}w_j^{t\arg et}}+(1-\lambda)\times\sum_{i=1}^{n-1}\frac{\sum_{j=1}^{l}w_j^{join}c_j^{join}(u_i,u_{i+1})}{\sum_{j=1}^{l}w_j^{join}} \quad (1)$$

其中,u_i 为长度为 n 的候选基元,t_i 为其目标基元,匹配代价包含 k 维的韵律特征子代价,拼接代价包含 l 维的声学特征子代价.每一维子代价的权重为 w_j 且由人工经验指定.目前也有一些基于机器学习的训练方法和基于人类感知的误差反向传播(error back propagation,简称 BP)算法来调整权重,取得了较好结果.λ为匹配代价和拼接代价的权重比值,通常设为 0.5,但是如果语料库录制效果欠佳,为了减小拼接失真,可以适度增大拼接代价比例.

2.2 特征选取

对于 diphone 基元,其匹配代价定义为一系列上下文韵律特征匹配程度之和来保证所选候选基元符合目标基元的上下文发音特征.本文采用的韵律特征见表 2[14].

Table 2 Prosodic features

表 2 韵律特征参数

特征	描述
短语	当前 diphone 在韵律短语中的位置
重音	当前 diphone 所在音节是否为重音音节
POS	Part of Speech,功能词或者文本词词性
音节	当前 diphone 在音节中的位置
单词	当前 diphone 在单词中的位置
前音素	前音素是否匹配
后音素	后音素是否匹配
小基元开销	同等韵律环境,优先选取较大基元
词典外词开销	优选选择符合 cmudict 发音词典的基元

而对于单词以及音节基元,为了简化模型,我们认为不同单词或者音节内部的韵律不匹配较为少见,或者即使某几个内部音素不匹配其协同发音也不会直接对拼接点造成严重影响,所以在本文中我们只考虑不同基元的候选基元内部首尾 diphone 的匹配代价,然后将首尾 diphone 的匹配代价之和作为整个单词或者音节的匹配代价.

对于拼接代价,所有基元类型都统一为首尾拼接点的声学特征曼哈顿距离之和.目前本系统中应用的拼接

代价声学特征主要有频谱信息(12 维 MFCC 系数及其一阶二阶差分)、激励信息(短时能量及其一阶二阶差分)、基频信息(*F*0,清音为–1)组成的 40 维特征参数.其中,所有特征的特征值均归一化到(0,1)之间的浮点值且总代价也进行了归一化处理.

最后,总的代价函数对应到具体不同层级的 Viterbi 算法中的状态值和转移值,将会在下一节描述.

2.3 不定长基元选取算法

本文提出的不定长选音算法实际上是一个分层的 Viterbi 搜索图.对于一段输入文本,前端文本分析阶段会生成目标基元序列,其候选基元构成一个二维网格拓扑结构,经典的 Viterbi 算法被用来选取一条最优路径使得基元序列的全局总代价最小.

寻找最佳基元搜索路径示例如图 3 所示.图 3 描述了由单词、音节、diphone 这 3 个逻辑层级组成的动态规划搜索路径.

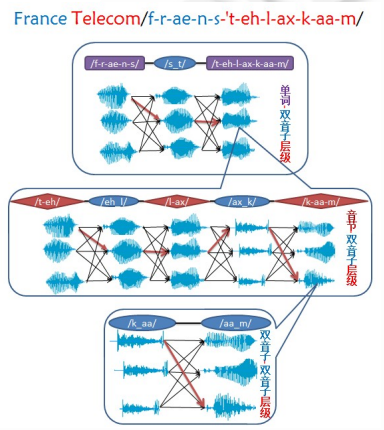

Fig.3 Example of multi-layer viterbi based non-uniform unit selection algorithm

图 3 基于分层 Viterbi 搜索的不定长基元选取算法示意图

在实际应用中,对于文本分析后的待合成目标序列,本文提出的具体算法如下:

1) 将待合成句子拆分为单词序列,单词边界采用第 1.3 节中描述方法修改为稳定段边界.然后在单词之间插入 diphone 来过渡前后两个单词的稳定段,句首和句末插入短暂静音过度,最后形成最初的目的基元序列.

2) 对于目的基元中的每个元素,根据具体基元的匹配代价从语料库中直接选取 *N*-best 候选队列,此外,设置每个基元的最大候选基元数 *M*,如果直接候选基元数 *N*<*M*,则将其拆分为其更低层级的基元序列,递归的使用 Viterbi 算法选取(*M*–*N*)个子单元序列,将这些子单元序列拼接为一个完整的上层基元,其匹配代价为全部子序列匹配代价之和,拼接代价则取首尾基元分别进行计算,然后和直接选取基元一起用 Viterbi 算法选取最优序列作为候选合成序列.

3) 其中,对于非直接选取基元的拆分,规则如下:

   a) 对于单词,如果其为单音单词(如单词 a,I 等),因为其发音特征已经包含在前后插入的 diphone 过渡基元中,故跳过处理;对于其他单词,则根据当前单词的音节组成结构,将其拆分为稳定切分边界的音节序列,然后在音节中间插入 diphone 进行稳定段过渡.通过拆分,对于非直接选取单词的选取则转变为对其音节子序列的最优路径序列的选取.

   b) 对于音节,类似地将非单音音节拆分为 diphone 序列,然后对其进行最优路径的选取.

   c) 对于 diphone,因为其已是本文所采用的最小合成基元,已不能拆分,所以,diphone 通过韵律匹配代价直接从语料库中选取而得.对于语料库中不存在的少部分低频 diphone,则采取声学特征相

似的 diphone 替换列表处理为语料库中已经存在的基元.通过替换,可以保证所有 diphone,哪怕极为稀疏的 diphone 都能找到符合要求的语料库样本,从而构成一个可以合成任何英语发音序列的鲁棒 TTS 系统.

3 实验结果与分析

本文采用 10 000 句大约 9 小时时长的大规模语料构建语音库,采用 CMU_ARCTIC[15]的前 50 句文本作为测试语料,随机选取 10 名非语音专长的测听者,然后通过平均主观意见打分(mean opinion score,简称 MOS)和听写测试来评测合成语音音质.

MOS 要求测听者听完随机打乱的包括自然语音、Festival 合成系统的标准基元选取合成语音以及本文提出的不定长选音方法合成语音,然后根据自然度、清晰度、可懂度、连续性等的好坏进行 1~5 分的打分.MOS 结果如图 4 所示.

听写测试则让测听者听完随机打乱的语音,然后写出他们听到的内容.其平均单词听写错误率(word error rate,简称 WER)如图 5 所示.

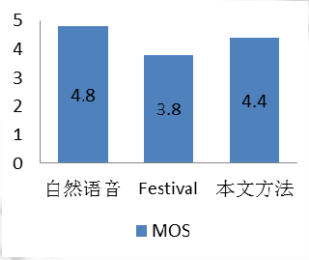

Fig.4 Mean opinion score

图 4 平均主观意见打分

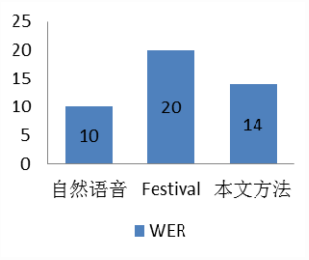

Fig.5 Word error rate

图 5 平均单词听写错误率

评测结果显示,虽然本文方法合成语音与自然语音相比仍有差距,但是相对于传统的基于 diphone 的基元选取拼接方法,本文方法合成的语音具有更高的自然度和音质.

4 结束语

本文在传统的基于代价函数的基元选取算法基础上,将 diphone 的稳定段边界模型应用到单词和音节中,最后使用 3 种基元模型的分层不定长选音算法从语料库中优选最佳合成基元序列拼接合成最终语音.该算法一方面利用分层统一的不定长选音策略,尽可能地选取具有更好韵律特性和声学连续性的较大基元,从而显著减少拼接点,将有可能发生协同发音或者切分错误的拼接点包含到更大的基元内部;另一方面通过稳定段切分修改传统拼接基元边界类型,充分利用了 diphone 良好的拼接特性,从而提高合成语音的连续性和自然度.评测结果显示,这种方法与传统 diphone 拼接合成方法相比,其合成效果有显著的提升.

References:

[1] Hunt AJ, Black AW. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of the ICASSP. 1996. 373−376.

[2] 倪昕.语料库支持的英语文语转换合成引擎[硕士学位论文].北京:清华大学,2004.

[3] 裴定瑜.基于大语料库英文 TTS 语音拼接单元的选择[硕士学位论文].上海:同济大学,2006.

[4] Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. of the ICASSP, Vol.3. 2001. 1315−1318.

[5] 胡克,康世胤,郝军.中文 HMM 参数化语音合成系统构建.通信技术,2012,45(8):101−103,108. [doi: 10.3969/j.issn.1002-0802.2012.08.032]

[6] Yamagishi J. An introduction to HMM-based speech synthesis. Technical Report, Tokyo Institute of Technology, 2006.

[7] Kang S, Qian X, Meng H. Multi-Distribution deep belief network for speech synthesis. In: Proc. of the ICASSP2013. 2013.

[8] Black AW, Taylor P. Automatically clustering similar units for unit selection in speech synthesis. In: Proc. of the EUROSPEECH, Vol.2. 1997. 601−604.

[9] Kishore SP, Black AW. Unit size in unit selection speech synthesis. In: Proc. of the INTERSPEECH. 2003.

[10] Latacz L, Kong YO, Verhelst W. Unit selection synthesis using long non-uniform units and phonemic identity matching. In: Proc. of the Blizzard Challenge Workshop. 2007.

[11] Chu M, Peng H, Yang HY, Chang E. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. In: Proc. of the ICASSP, Vol.2. 2001. 785−788.

[12] Black AW, Clark R. The festival speech synthesis system. 2011. http://www.cstr.ed.ac.uk/projects/festival/

[13] 倪昕,蔡莲红.基于混合基元模型的非定长基元选取算法.小型微型计算机系统,2005,6:1079−1082.

[14] Clark RAJ, Richmond K, King S. Multisyn: Open-Domain unit selection for the Festival speech synthesis system. Speech Communication, 2007,49(4):317−330.

[15] Kominek J, Black AW. The CMU Arctic speech databases. In: Proc. of the 5th ISCA Workshop on Speech Synthesis. 2004.

王欣(1991－),男,甘肃陇西人,硕士,主要研究领域为语音合成.
E-mail: xin-wang11@mails.tsinghua.edu.cn

蔡莲红(1945－),女,教授,博士生导师,主要研究领域为人机语音交互,语音合成,多媒体技术.
E-mail: clh-dcs@tsinghua.edu.cn

吴志勇(1977－),男,博士,副研究员,博士生导师,主要研究领域为语音处理,语音合成.
E-mail: zywu@sz.tsinghua.edu.cn