ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.5, May 2006, pp.983−990
DOI: 10.1360/jos170983

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

基于虚拟不定长的语音库裁剪方法*

张 巍[1+], 吴晓如[2], 赵志伟[2], 王仁华[1]

[1](中国科学技术大学 电子工程与信息科学系,安徽 合肥 230027)

[2](安徽中科大讯飞信息科技有限公司,安徽 合肥 230088)

Virtual Non-Uniform Synthesis Instances Pruning Approach for Corpus-Based Speech Synthesis System

ZHANG Wei[1+], WU Xiao-Ru[2], ZHAO Zhi-Wei[2], WANG Ren-Hua[1]

[1](Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China)

[2](Anhui USTC Iflytek Co., Ltd., Hefei 230088, China)

+ Corresponding author: Phn: +86-551-5331851 ext 8048, E-mail: weizhang@ustc.edu.cn, http://cheung.colin.googlepages.com

Zhang W, Wu XR, Zhao ZW, Wang RH. Virtual non-uniform synthesis instances pruning approach for corpus-based speech synthesis system. *Journal of Software*, 2006,17(5):983−990. http://www.jos.org.cn/1000-9825/17/983.htm

Abstract: Tailoring voice font, or pruning redundant synthesis instances, is an important issue of scalable Corpus-based Text To Speech (TTS) system. However, pruning redundant synthesis instances, usually results in the loss of non-uniform. In order to solve this problem, the concept of virtual non-uniform is proposed. According to this concept and the synthesis frequency of each instance, an algorithm named StaRp-VPA is constructed to make TTS scalable to hardware. In experiments, the naturalness scored by Mean Opinion Score (MOS) remains almost unchanged when less than 50% instances are pruned off, and the MOS does not severely degrade when the reduction rate is above 50%.

Key words: speech synthesis; text to speech; pruning redundant synthesis instances; scalable speech synthesis system

摘 要: 语音库裁剪或语音库去冗余,是大语料库语音合成技术的一个重要问题.提出了虚拟不定长替换的概念,以弥补不定长的损失.结合合成使用变体的频度,构建了语音库裁剪算法 StaRp-VPA.该算法能够以任意比例裁剪语音库.实验表明:当裁剪率小于50%时,合成自然度几乎没有下降;当裁剪率大于50%时,合成自然度也不会严重降低.

关键词: 语音合成;文语转换;语音库裁剪;可伸缩语音合成系统

中图法分类号: TP391 文献标识码: A

基于单元选择(unit selection)的语音合成技术由于采用真人的发音片段作为语音合成的单元,所以能够产生很高音质的合成语音,是目前应用比较成功的语音合成方法.在此基础上,为了提高语音合成的自然度和可懂

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2004AA114030 (国家高技术研究发展计划(863))

Received 2005-05-20; Accepted 2005-10-10

度,发展出了基于语料库的语音合成(corpus-based TTS(text to speech)),代表了语音合成的最高水平[1−4].对于大语料库语音合成而言,合成音质不再是主要问题,而合成的自然度和可懂度是衡量其合成质量最重要的指标(本文提到的合成质量均以自然度和可懂度为指标).

这种合成方法将数据挖掘和知识发现领域兴起的数据驱动技术和数字信号处理技术有机地结合在一起.合成系统 KBCE*采用的也是这种方法,其合成汉语时,最小单位为音节(syllable).这些音节一般利用 Viterbi[5]算法从语音库中挑出(selection).语音库中包含录制好的语音和索引.索引的基本单位是语音单元和声学变体**.一个语音单元(unit)可以是单音节和不定长(non-uniform unit,若干个连续的单音节组成).每个单元按照不同的高层韵律环境和声学特征,包含许多不同的声学变体(variant,instance 或称 font).实际上,单元是一个索引树,称为不定长分类树(一般通过基于问题集的分类或聚类 CART[6]方法构建),其所含的变体隶属于不同的叶子节点,如图 1 所示.

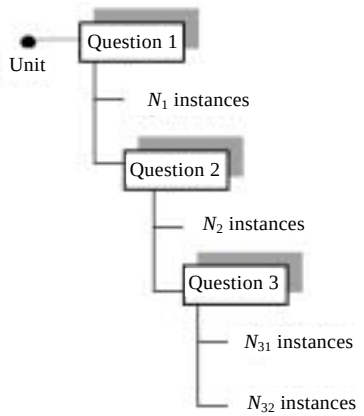

Fig.1 Unit tree and Instances
图 1 单元树和变体示意

在这一类方法中,语音合成问题就转化为对语音库获取、标注、索引和搜索[1−4].为了得到自然的合成语音,往往需要大量变体(在 KBCE 使用的 GB 级音库中,会有几个至十几个小时的语音).在这样超大规模的音库中进行合成所必需的存储、加载和搜索比较耗时,因此,大语料合成系统对硬件的要求较高.如果能在保证合成质量的前提下适当减小语音库,将使得大语料库合成方法具有更好的适应性;如果能更进一步,在任何应用环境下给出大小合适的音库,将使得语料库合成方法具有可伸缩性(scalability).这些都涉及到语音库去冗余或称语音库的裁剪问题.

其实,音库存在一定的冗余,例如:一个单元的某些变体合成系统几乎不会使用,一些变体甚至是发音不够理想的孤立点;一个单元有些变体可以相互替换.前人已经提出一些语音库裁剪(或称音库去冗余)的方法.文献[7]以双音素(diphone)为单位,进行基于声韵问题集的聚类,离聚类中心较远的变体被裁剪.这种方法保留 50%以上的变体时,合成质量不会有很严重的下降.文献[8]以马尔可夫模型进行裁剪——HMM 得分最高的那些变体被作为类中心保留下来.文献[9,10]中提出了赋权矢量量化(WVQ)的方法进行裁剪,音库裁剪最大到 50%时也不会产生严重失真.文献[11]统计合成系统使用每个双音素的频率,利用数据库缩减技术进行裁剪.实验结果也显示保留率在 50%以上,合成质量不会有严重下降.文献[12]中提出韵律孤立点和变体重要性(对合成的贡献)的概念,并以此出发进行裁剪.该方法在音库裁剪到 50%时,合成质量几乎没有下降.我们针对嵌入式应用,采用模式聚类方法,进行了音库裁剪和压缩的一些研究[13].

不定长技术保证是语料库合成语音高自然度、高可懂度的关键,而裁剪往往会导致不定长破坏和损失.针对这一问题,本文提出虚拟不定长(virtual non-uniform)替换的概念,意图尽量减小和弥补不定长的损失,并结合合成使用变体的频度,给出了一种语音库裁剪的新方法.这种方法在 KBCE 系统上得以实现.给出了 StaRp-VPA 算法,该算法可以任意比例裁剪语音库.大规模测听表明:当裁剪率小于 50%时,合成自然度几乎没有下降;当裁剪率大于 50%时,合成自然度也不会严重降低(详见第 3 节).

本文第 1 节从方法的思想入手,分析裁剪的关键技术,并提出虚拟不定长的概念.第 2 节给出 StaRp-VPA 算法的形式化描述.第 3 节给出详细的主、客观评测结果.第 4 节作总结分析.

* KBCE 是中科大讯飞信息科技有限公司多个核心语音产品的原型系统,在 1998 年、2003 年和 2004 年国家高技术研究发展计划(863)语音合成评比中均获第一名.

** 在本文中,语音单元表示一个汉语的字、词或共现连续字,不计韵律环境.声学变体表示语音单元在不同韵律和声学环境下的发音个体.

1 StaRp-VPA 算法的思想和裁剪的关键技术

语音单元一般都是在汉语中的高频字词或者共现连续字,冗余的可能性非常小[3].裁剪掉语音单元意味着某种韵律和声学环境的缺失,合成效果会受到较大影响.其实,冗余主要是由语音单元中多余变体引起的.这些多余的变体很少被合成使用.另外,有些变体之间差异性非常小,它们之间可以相互替代,保留其中之一即可.因此,语音库的冗余一般是由于变体过多.裁剪掉冗余的变体就是本文的出发点.

冗余变体的裁剪涉及到两个问题:(1) 每个语音单元中有多少比例的变体是冗余的;(2) 每个语音单元中,哪些变体是冗余的.也即在每个语音单元中,怎样确定各变体之间相对的重要程度.

对于问题(1),由于目前语料库都采用类似文献[3]的设计方式,可以认为,如果一个语音单元包含的声学变体越多,那么它含有的冗余变体就越多.因此,可以采用变率裁剪的方法,即给定总体保留率,在局部自动调整:一个单元的保留率与其所含变体数量相关:含有较多声学变体的保留率较小;含有较少变体的保留率较大.然后,对过小的保留率进行适当补偿;对于问题(2),由本文前面的讨论可以看出,衡量变体重要性的指标有两个:一是这个声学变体(实际上是它所代表的韵律环境)用于合成的频繁程度,被频繁使用的变体应该保留;另一个是这个声学变体可以用于代替其他声学变体的能力,替换能力越高的变体越应该保留.这个替换应该包括所有的不定长和单音节变体.变体重要性度量应该是这两个指标的函数(一个合理的形式是乘积函数),本文称其为变体的打分函数.单元变体按照分数值进行排序,排在后面的即为冗余变体.以上分析给出了本文提出的裁剪方法的基本思想.

根据这一思想,裁剪中有以下 4 个关键技术点:

(1) 变率裁剪中保留率的计算.若要将语音库裁剪到原来的$\beta(0<\beta\le 1)$,由问题 1 的分析,需要对不同单元 i 的变体采用不同的保留率 g_i(设其裁剪率为 t_i,则 $t_i=1-g_i$),同时使得整体的保留率仍然不变.设每个单元的变体占所有变体的比例为 p_i,则有

$$\sum_{i=1}^{I} p_i g_i = \beta \cdot$$

令 $p_i g_i=\beta/I$,则每个

$$g_i=\beta/I/p_i \tag{1}$$

可以看出,g_i 和 p_i 成反比,说明单元中变体越多,裁剪的越多,gn 问题(1)的分析相吻合.对于求出的 $\beta/I/p_i>1$ 的那些单元,$g_i=1$(就是完全保留),此时残差为

$$\sigma=\sum_{i=1}^{I}\left(\frac{\beta}{I}-p_i g_i\right)=\sum_{g_i=1}\left(\frac{\beta}{I}-p_i\right)+\sum_{g_i<1}\left(\frac{\beta}{I}-p_i g_i\right) \tag{2}$$

期望概率 Efr_i 描述一般文本中单元 i 出现的概率,其真实值可以通过大语料统计进行很好的近似(本文通过统计 300M 各种类型的文本得到);当前频率比 Sfr_i 描述在当前语音库中,单元 i 频率占所有单元频率的比例,其值可以通过对语音库变体数统计得到(Sfr_i 值随着保留率的补偿而不断更新).

$$x_i=Efr_i/Sfr_i,$$

描述当前频率比和期望概率的差距,用 x_i 来进行保留率计算的残差补偿.

$$G_i=g_i+\sigma\frac{x_i}{\sum_{g_i<1}x_i}=\frac{\beta}{Ip_i}+\sigma\frac{x_i}{\sum_{g_i<1}x_i},\text{对于所有 } g_i<1 \tag{3}$$

此时,如果 $G_i\le 1$,那么 $g_i=G_i$;如果存在 $G_i>1$,则令 $g_i=1$.补偿过程反复进行,当计算出的所有保留率均小于或等于 1 时终止.保留率补偿是为了使变体多的单元中不会过分裁剪,从而尽量保证韵律和声学环境的完整性.

(2) 虚拟不定长和变体打分(instance importance scoring).在变体得分中包括两个主要参数:使用系数和替换得分.一个声学变体 L 的使用系数定义为:去除该变体所在索引树的叶子、语料库动态覆盖率的损失比.

变体 L 去除前,语料库覆盖率(计算方法见文献[3])为 A_{0L},去除后,语料库覆盖率为 A_L,则该声学变体的使用系数为

$$\alpha_L=(A_{0L}-A_L)/A_{0L}.$$

使用系数是从语料层来衡量变体的重要性.在同一个韵律环境和声学环境下,各变体使用系数相同;在不同韵律和声学环境下,各变体使用系数一般不同.

语音库裁剪会导致不定长的损失.为了尽可能减小不定长的损失,这里引入虚拟不定长的概念.

对于某个变体,在语音库中去除这个变体(这个变体本身称为第 0 替换 R_0),由合成系统根据某种度量(本文使用的是 Viterbi,除此之外还可以使用声学距离或 Trainable 等方法)挑选出第 1 替换 R_1.R_1 不是真实的不定长,而是真实不定长 R_0 的当前最佳替换.为了达到最佳目的,挑选需要完全忠实于 R_0 的高层韵律环境,我们形象地称其为“语音填空”.

同理,在语音库中去除第 0 和第 1 替换,由合成系统挑选 R_0 的第 2 替换.依此类推,去除前 0,…,N–1 个替换,可以得到不定长的第 N 个替换 R_N 来替换 R_i(0<i<N),称为变体 R_0 的虚拟不定长.

挑选虚拟不定长的度量方法(本文为 Viterbi 方法,KBCE 的 Viterbi 计算分为两个部分:一部分称为自身代价——表征了预选变体和目标的差异,是高层韵律环境、基频及时长的函数;另一部分为连接代价,它表征相互连接的两个不定长之间的基频差异).给出每个 R_K 的代价 Q_K,它正好描述了虚拟不定长与真实不定长的差异,并且满足单调性:Q_0=0,$Q_{K-1}\le Q_K$.

各替换得分定义为

$$M_K=\exp\left(-\frac{Q_K^2}{\sigma}\right),$$

其中σ为宽度,用于控制可以响应的 Q_K 范围.原始不定长的 M_0=1,由 Q_K 的单调性,可以看出 M_k 满足单调性,即 $M_K\le M_{K-1}<1=M_0$.

语音库的裁剪最终是对单音节变体的裁剪,因此必须将替换得分加到组成替换的单音节变体中去.例如,对于变体“中国人”的某个替换(设其替换得分为 M_F),它由单音节变体“中”、“国”和“人”构成,对于这些单音变体都要进行加分:$\alpha_F M_F$,其中α_F 为真实不定长“中国人”的使用系数.

一个单音节变体 m 的总得分 S_m 如下:

$$S_m=\sum_{j\ge L_m}^{J}\eta_j mark_j,\quad mark_j=\sum_{n=0}^{N}F_n,\quad F_n=\alpha_n M_n,$$

其中:M_n 是变体 m 参与的音节数为 j 的不定长(包括*真实*和虚拟)替换得分;α_n 是该不定长变体的使用系数;F_n 就是一种加权替换得分.因此,$mark_j$ 是变体 m 所能参与的所有长度为 j 的不定长替换总得分;L_m 为变体 m 的音节数.将不同长度不定长替换得分求和就得到了 S_m.同时,为了平衡各种长度变体之间相对的重要程度,加权系数 η_j 一般为 1.可以看到,S_m 中考虑到了每个变体替换其他变体的能力(替换得分),这种替换同时考虑到了每个变体被合成系统使用的频繁程度(使用系数);如果以音节数为粒度标准,S_m 考虑了同粒度替换,也考虑了不同粒度的替换.

(3) 不定长变体调整.当裁剪掉单音节变体时,由它构成的不定长变体会发生损失.因此,每个变体上都要记录 N 个替换的信息.当裁剪完成后,如果不定长变体 R_0 的替换 R_K 中有一个参与的单音节被裁剪掉,则这个替换 R_K 也被裁剪掉.剩下的编号 L 最小的替换 R_L 可以作为 R_0 的替换,最终在不定长变体索引中用 R_L 代替 R_0.这种方法减小了裁剪对不定长的破坏:分值较高的 R_0 会被保留下来,此时 R_L=R_0,不定长不发生损失;如果 R_0 被裁剪,可以由虚拟不定长 R_L 代替,由于 R_L 没有 R_0 好,不定长部分损失;只有在所有的替换均被裁剪时,不定长才会完全损失.由此可见,替换数 N 必须加以合理设置:如果设置太小,则替换有可能都被裁剪掉,完全损失的不定长较多;如果设置太大,则实际算法的空间和时间开销较大.本文中,N=5.

(4) 联动打分的消解.在给每个变体打分时,存在这样的问题:可以替代的两个声学变体相互重复加分——本文称其为联动打分.举例来说:对于某两个可以相互替代的变体 V_1 和 V_2,在对所有 V_1 的替换打分时,V_2 被加分;而对 V_2 的替换打分时,V_1 被加分.由于 V_1 和 V_2 可以相互替换,因此取其中之一即可,而此处它们却被重复加分,可能导致最终都被保留.对于这一点,打分的方法不必改变,只要采用如下方法就可以消解(本文称其为联动打

分的消解(associated-scoring elimination)).

每个变体 V 有两种结构:可以替换 V 的变体组成 $V.REL$,V 可以替换的变体组成 $V.RIL$.变体 V 的得分为

$$S_V=\sum_{R\in V.RIL} score_R .$$

假设分值排序为 V_1 和 $V_2,\ldots,V_H$,按照保留率只能取其中的 $k(k<H)$个.首先保留变体 V_1,因其分值最高;对于 $\forall R_x\in V_1.REL, R_x\in V_1.REL\Leftrightarrow V_1\in R_x.RIL$,从 $R_x.RIL$ 中删除 V_1,然后对 $R_x(R_x=V_2,\ldots,V_H)$进行分值调整

$$S_{R_x}=S_{R_X}-score_{V_1} .$$

最后,重新排序 $V_2,\ldots,V_H$,取分值最大的变体,依此进行下去.

如图 2 所示,如果 V_2 和 V_1 之间可以替换,那么有:(a) $V_1\in V_2.REL\Leftrightarrow V_2\in V_1.RIL$;(b) $V_2\in V_1.REL\Leftrightarrow V_1\in V_2.RIL$,则由上面过程,从 $V_2.RIL$ 中删除 V_1,消除了(b),就不会再产生 V_1 和 V_2 重复计分的情况,联动打分被消解.联动打分的消解使得语音库裁剪化归为:首先构建赋权有向图,然后从出度最大的节点开始,寻找图中那些出度最大而入度最小的节点[14].

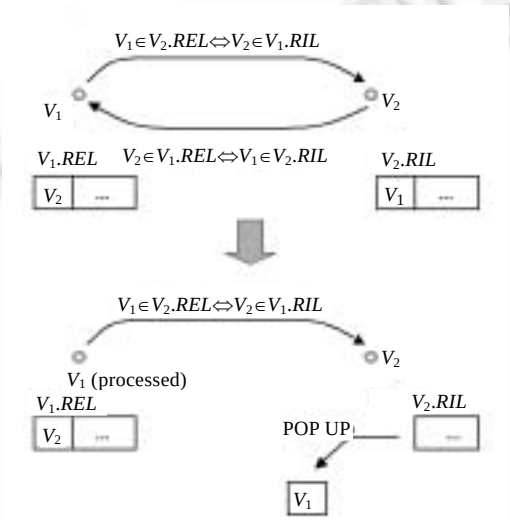

Fig.2 Associated-Scoring elimination

图 2 联动打分的消解

2 StaRp-VPA 算法的形式化描述

本节给出 Statistics & Replacing based Variant Pruning Algorithm 算法的形式化描述(StaRp-VPA).

Step 1: Computing the instances reserve rate of each unit.

Compute from equation (1) to equation (3).

Step 2: IIS scoring of instances.

(1) $\forall V$, $Length(V)=L$, execute (1.1) and (1.2):

(1.1) $V.REL=\{R_0,R_1,\ldots,R_N\}$, $scores=F_0,F_1,\ldots,F_N$.

(1.2) $\forall W\in V.REL$, $W.RIL=\{V\}\cup W.RIL$.

(2) $\forall U$, execute from (2.1) to (2.2):

(2.1) $Reserve\_Num=Number(U)\times Reserve\_rate(V)$

(2.2) $\forall V\in U$, execute (2.2.1):

(2.2.1) For i=1 to $Reserve\_Num$,

Execute ASE: Associated-Scoring Elimination

(2.2.2) Tailor all Variants left by ASE

Note: L is syllable number of instance; its maximum is max length.

V=Variant represents instances; U=Unit represents units.

Number(U) is number of instances that unit U includes.

Step 3: Adjusting speech database.

3 主客观评测结果

StaRp-VPA 算法在 KBCE 合成系统上分别给出了各种裁剪比例的语音库.本节将给出这些语音库合成效果的主客观评测.

3.1 客观度量

StaRp-VPA 算法的主要目的是尽量减小不定长的损失,这里以裁剪后不定长(包括单音节)的分布情况为客观度量.包括:裁剪后的语音库中,保留的不定长数和原始库的不定长数之比(*rONU*),虚拟不定长数和原始库的不定长数之比(*rVNU*),被裁剪的不定长数和原始库的不定长数之比(*rTNU*),以及下面将要说明的λ_O 和λ_{OV}.表 1 给出了语音库在各保留率下,不定长的分布情况.其中,β为语音库的保留率,语音库的裁剪率为 1–β.

$$\lambda_O = \frac{rONU}{\beta}, \lambda_{OV} = \frac{rONU + rVNU}{\beta}.$$

Table 1 Non-Uniform distribution with different reserve rate

表 **1** 各保留率的语音库中不定长分布

β (%)	*RONU* (%)	*RVNU* (%)	*RTNU* (%)	λ_O	λ_{OV}
73	62.53	36.43	1.24	0.86	1.6
61.9	47.80	48.85	3.35	0.77	1.56
50	33.98	56.44	9.58	0.68	1.81
30	15.93	48.22	35.85	0.53	2.14
10	4.36	19.15	76.49	0.43	2.35

本文称λ_O 为原始不定长保留比,称λ_{OV} 为不定长保留比.它们描述了保留率下降对不定长损失的影响.一般来说,语音库裁剪后,不定长数量一定会发生减少.最理想的情况是β=*rONU*,但这几乎是不可能的.因为对于某个单音节变体 V,会有 L_V 个包含不同音节数的不定长变体与其相关联,当 V 被裁剪时,L_V 个不定长都会损失;V 不同,L_V 也不同.如果 L_V 值较大的变体 V 被裁剪,*rONU*<<β.从表 1 和图 3 中可以看出λ_O 下降的速度比较慢,这说明 StaRp-VPA 算法倾向于保留那些 L_V 值较大的变体.通过虚拟不定长替换可以在一定程度上弥补原始不定长损失,λ_{OV} 正好反映了这一情况.从表 1 和图 3 中可以看出,由于采用了虚拟不定长,使得λ_{OV}>1.

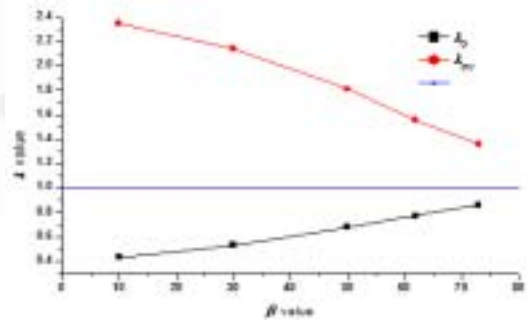

Fig.3 Distribution of β, λ_O and λ_{OV} with different reserve rate

图 3 不同保留率β的λ_O 和λ_{OV} 分布

3.2 主观度量

本文使用两种类型的语料,对不同规模裁剪库的合成效果进行了主观测听.主观度量采用 MOS(mean opinion score)分.语料 1 共有 150 句语料,通过使用文献[3]的语料搜索方法得到,前 100 句为覆盖率最高的,后 50

句为覆盖率较低的.由 5 位测听员对不同裁剪库分别进行了测听.表 2 为前 100 句的 MOS 分,表 3 为后 50 句的 MOS 分.可以看出,当裁剪率较高时,MOS 下降得并不多;即使在语音库只保留 30%时,MOS 降低也没有超过 0.07;特别是对于后 50 句,有 73%甚至比原始库的合成效果还要稍好.

Table 2 MOS of front 100 sentences in Corpus 1

表 **2** 语料 1 前 100 句的 MOS 分

Reserve rate (%)	Listener *A*	Listener *B*	Listener *C*	Listener *D*	Listener *E*	MOS
30	3.83	3.63	3.6	3.51	3.77	3.668
50	3.85	3.69	3.61	3.52	3.83	3.7
62	3.85	3.69	3.62	3.54	3.86	3.712
73	3.88	3.71	3.65	3.54	3.88	3.732
100	3.87	3.7	3.64	3.54	3.93	3.736

Table 3 MOS of rear 50 sentences in Corpus 1

表 **3** 语料 1 后 50 句的 MOS 分

Reserve rate (%)	Listener *A*	Listener *B*	Listener *C*	Listener *D*	Listener *E*	MOS
30	3.83	3.69	3.51	3.4	3.72	3.63
50	3.86	3.76	3.57	3.44	3.78	3.682
62	3.86	3.79	3.55	3.43	3.83	3.692
73	3.85	3.79	3.58	3.42	3.85	3.698
100	3.85	3.77	3.58	3.45	3.83	3.696

语料 2 共有 100 句语料,为从网页上抓取的覆盖率最高的新文本 100 句.由另外 5 位测听员对不同裁剪库分别进行了测听.表 4 为这 100 句的 MOS 分.可以看出,语料 2 的 MOS 分下降也不多,只是比语料 1 中稍多一点.这里甚至将语音库只保留 10%,此时 MOS 下降了 0.22.

Table 4 MOS of Corpus 2

表 **4** 语料 2 的 MOS 分

Reserve rate (%)	Listener 1	Listener 2	Listener 3	Listener 4	Listener 5	MOS
10	3.5	3.83	3.45	3.73	3.09	3.52
30	3.61	3.87	3.57	3.9	3.26	3.642
50	3.6	3.89	3.62	3.89	3.3	3.66
73	3.72	3.9	3.69	4	3.32	3.726
100	3.69	3.93	3.73	4	3.35	3.74

由图 4 可以看出,MOS 分曲线下降得比较缓慢(注意,图 4 中 Y 轴标度在 3.0~3.8 之间):当保留率大于 50%时,MOS 分几乎没有下降;当保留率小于 50%时,MOS 分也不会严重降低.

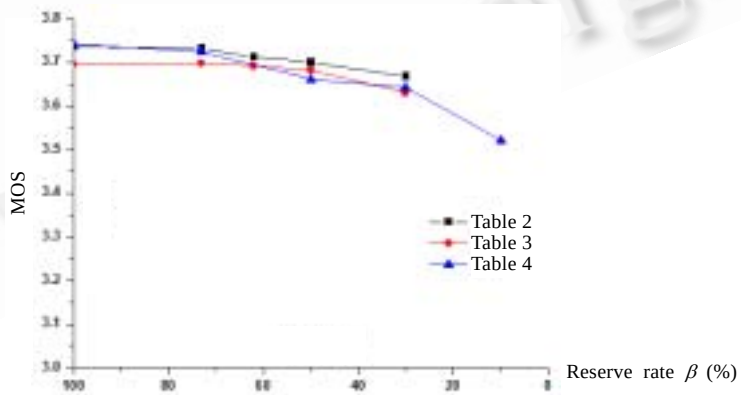

Fig.4 The change of MOS with different reserve rate

图 4 两个语料中,MOS 随不同保留率的变化

4 分析与结论

在测听实验中,MOS 下降得比较慢,经分析可能是因为:(1) StaRp-VPA 算法的机制,替换能力强和使用频繁

的不定长被保留下来;(2) 对于被替换掉的不定长,虚拟不定长在一定程度上弥补了原始不定长的损失;(3) 裁剪率自动调整,使得单元韵律环境得以保留,而只是不重要的变体被裁剪.

本文提出虚拟不定长(virtual non-uniform)的替换概念,在一定程度上弥补了不定长的损失.以此结合变体使用系数给出的裁剪算法 StaRp-VPA,在 KBCE 系统上得到各种裁剪率的语音库.测听表明:当裁剪率小于 50%时,MOS 分几乎没有下降;当裁剪率大于 50%时,MOS 分也不会严重降低.

致谢 感谢刘庆峰博士和胡国平博士对项目的支持.感谢张冰硕士对英文摘要和全文文字的润色.

References:

[1] Hunt A, Black A. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of the ICASSP'96, Vol.1. 1996. 373−376.

[2] Sagisaka Y, Kaiki N, Iwahashi N, Mimura K. ATR-v-TALK speech synthesis system. In: Proc. of the ICSLP'92, Vol.1. 1992. 483−486.

[3] Liu QF. Speech synthesis study based on perception quantification [Ph.D. Thesis]. Hefei: University of Science and Technology of China, 2003 (in Chinese with English abstract).

[4] Chu M, Peng H, Yang H, Chang E. Selection non-uniform units from a very large corpus for concatenative speech synthesizer. In: Proc. of the ICASSP 2001. 2001.

[5] Rabiner LR. A tutorial on hidden markov models and selected application in speech recognition. Proc. of the IEEE, 1989,77(2): 257−285.

[6] Breiman L, Friedman J, Olsen R, Stone C. Classification and regression trees. Pacific Grove: Wadsworth and Brooks, 1984.

[7] Black AW, Taylor PA. Automatically clustering similar units for units selection in speech synthesis. In: Proc. of the Eurospeech'97, Vol.2. 1997. 601−604.

[8] Hon H, Acero A, Huang X, Liu J, Plumpe M. Automatic generation of synthesis units for trainable text-to-speech systems. In: Proc. of the ICASSP '98, Vol.1. 1998. 293−296.

[9] Kim SH, Lee YL, Hirose K. Pruning of redundant synthesis instances based on weight vector quantization. In: Proc. of the Eurospeech 2001. 2001. 2231−2234.

[10] Kim SH, Lee YL, Hirose K. Unit generation based on phrase break strength and pruning for corpus-based text-to-speech. ETRI Journal, 2001,23(4):168−176.

[11] Rutten P, Aylett M, Fackrell J, Taylor P. A statistically motivated database pruning technique for unit selection synthesis. In: Proc. of the ICSLP 2002. 2002. 125−128.

[12] Zhao Y, Chu M, Peng H, Eric C. Custom-Tailoring TTS voice font-keeping the naturalness when reducing database size. In: Proc. of the Eurospeech 2003. 2003. 2957−2960.

[13] Ling ZH, Hu Y, Shuang ZW, Wang RH. Compression of speech database by feature separation and pattern clustering using STRAIGHT. In: Proc. of the ICSLP 2004. 2004. 766−769.

[14] Bondy JA, Murty USR. Graph theory with application. New York: American Elsevier, 1976.

附中文参考文献:

[3] 刘庆峰.基于听感量化的语音合成研究[博士学位论文].合肥:中国科学技术大学,2003.

张巍(1975－),男,北京人,博士,主要研究领域为数据挖掘和统计分析,数据驱动技术,可伸缩语音合成系统.

赵志伟(1978－),男,助理工程师,主要研究领域为大语料库语音合成.

吴晓如(1972－),男,博士,副高级工程师,主要研究领域为语音合成,语音识别.

王仁华(1943－),男,教授,博士生导师,主要研究领域为语音合成和语音识别,数字信号处理.