

语音库裁剪的一种不定长递阶聚类方法

张 巍^{1),2)} 吴晓如³⁾ 刘 江³⁾ 王仁华²⁾

¹⁾(中国海洋大学计算机科学系 山东 青岛 266100)

²⁾(中国科学技术大学电子工程与信息科学系 合肥 230027)

³⁾(安徽中科大讯飞信息科技有限公司 合肥 230088)

摘 要 大量使用不定长是大语料库语音合成质量的一个重要保证,而语音库裁剪方法通常会导致不定长的损失.针对这一关键性问题,该文构建了 NuClustering-VPA 算法:对不同粒度的不定长变体进行聚类,根据高阶聚类结果调整低阶变体的聚类,使得低阶聚类中心有所偏向. NuClustering-VPA 算法保留了最重要的不定长,从而有效减小了裁剪对不定长的破坏. 测听实验表明,利用 NuClustering-VPA 算法,即使在语音库裁减率为 39.63% 时,合成自然度下降较小,仍然保持在较高的水平. 这一技术已被应用在科大讯飞公司的实际语音产品中.

关键词 基于语料库的语音合成;语音库裁剪;语音库去冗余;可伸缩语音合成系统

中图法分类号 TP391

A Non-Uniform Clustering Synthesis Instances Pruning Approach for Corpus-Based TTS

ZHANG Wei^{1),2)} WU Xiao-Ru³⁾ LIU Jiang³⁾ WANG Ren-Hua²⁾

¹⁾(Department of Computer Science, Ocean University of China, Qingdao, Shandong 266100)

²⁾(Department of Electronic Engineering and Information Science, University of Science & Technology of China, Hefei 230027)

³⁾(Anhui USTC Iflytek Co., Ltd., Hefei 230088)

Abstract The employment of non-uniform does great help for Corpus-based TTS to synthesize high natural speech. But Tailoring TTS voice font, or pruning redundant synthesis instances, usually results in loss of non-uniform. In order to solve this problem, this paper proposes the algorithm named NuClustering-VPA. According to this algorithm, the high level non-uniforms containing same syllables are clustered to several centers, then the centers are projected to low level non-uniforms. Therefore, the center's projections can guide the clustering of low level non-uniforms. These series of processes avoid erasing or destroying those key non-uniforms for synthesis. In experiments, the naturalness scored by MOS does not severely degrade when reduction rate is above 39.63%. And this approach has been applied in software products of Iflytek Co. Ltd.

Keywords Corpus-based TTS; Tailoring TTS voice font; pruning redundant synthesis instances; scalable TTS

收稿日期:2005-05-30;最终修改稿收到日期:2007-03-20. 本课题得到国家自然科学基金(60602017)和国家“八六三”高技术研究发展计划项目基金(2004AA114030)资助. 张 巍,1975 年生,博士,副教授,硕士生导师,研究方向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. E-mail: weizhang@ouc.edu.cn. 吴晓如,1972 年生,博士,主要研究方向为语音合成和语音识别. 刘 江,1980 年生,学士,研究方向为大语料库语音合成. 王仁华,1943 年生,教授,博士生导师,研究领域为语音合成和语音识别、数字信号处理.

1 引 言

大语料库的语音合成(Corpus-based TTS),或称基于选择的语音合成(Selection-based TTS),能够产生高自然度的合成语音,是目前应用较多的语音合成方法^[1-4].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融汇在一起.

对于汉语来说,合成最小单位一般为音节(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:语音单元和声学变体^①.一个语音单元(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聚类 CART^[6]方法构建),其所含的变体隶属于不同的叶子节点.图 1 给出了一个单元和变体关系的示意图.

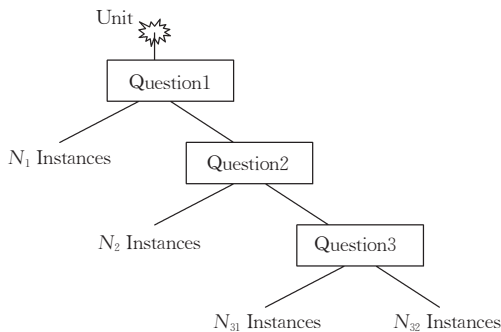


图 1 单元和变体示意

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(会有几个至十几个小时的语音).在这样超大规模的音库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,这涉及到语音库去冗余或称语音库的裁剪问题.

其实音库存在一定冗余:诸如,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,等等.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].

不定长的引入在基于语料库的语音合成技术中是一个非常重要的进步,目前的语料库语音合成引擎几乎都使用不定长技术^[1-4].各种粒度的不定长加大了语音库中变体与合成文本的匹配程度,一定程度上有效避免了合成时音节和音素之间的不连续和跳跃.从某种意义上说,不定长的质量直接决定了语料库语音合成的效果.而裁剪往往会导致不定长破坏和损失,目前提出的各种裁剪方法都还没有明确考虑这一问题,也没有比较好的解决方法.

怎样减小和避免裁剪导致的不定长破坏和损失——这就是本文研究的出发点.针对这一问题,本文提出一种 NuClustering-VPA 算法:对不同粒度的不定长变体进行递阶聚类,根据高阶聚类结果调整低阶变体的聚类,使得低阶聚类中心有所偏向. NuClustering-VPA 算法保留那些对于不同粒度不定长来说,在声韵上最为重要的变体,从而减小了裁剪对不定长的破坏. NuClustering-VPA 算法在 KBCE^②系统上得到了实现.测听实验表明,即使在语音库裁减率为 39.63% 时,合成自然度下降很小,仍然保持在较高的水平.科大讯飞公司利用这一技术,将面向企业级用户的超大规模语音库进行了适当裁剪,推出了面向桌面的语音系统——文语通,这一软件产品得到了用户的肯定.

本文第 2 节介绍递阶聚类裁剪方法的基本问题,

① 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.

② KBCE 是中科大讯飞信息科技有限公司语音产品 Interphonic 的原型系统.在具有高自然度合成语音的 KBCE 系统上,如果裁减后的合成语音仍具有较高自然度,那么裁剪的研究将具有重要的意义.

并介绍 NuClustering-VPA 的形式化描述;第 3 节详细介绍评测的结果和分析;第 4 节总结并指出下一步工作。

2 递阶聚类裁剪的基本问题和 NuClustering-VPA 算法

2.1 单元冗余和变体冗余

冗余可能发生在两个方面,一是语音单元出现冗余,二是单元的变体出现冗余。

语音单元一般都是在汉语中的高频字词,或者共现连续字,冗余的可能性不大^[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。因此只有包含变体过少的单元才会被裁剪。对于一个单元,设定 G_a 为单元保留门限,如果一个单元的变体数目小于 G_a ,该单元才被删除。本文称语音单元的裁剪为 URE(Unit Redundancy Elimination)。

冗余的另一个来源是语音单元中多余变体。这些多余的变体或者较少被合成使用;或者有些变体之间差异性非常小,他们之间可以相互替代,保留其一即可。

语音库的裁剪就是要剔除冗余的单元和单元中冗余的变体。这就是本文出发点。

2.2 绝对冗余、相对冗余和坏变体

对于变体,其冗余的程度是不相同的。本文通过分析,划归为三种类型:

绝对冗余指的是多个变体在韵律和声学上,差距非常小,在合成中,其所起到的作用是完全可以互换的。在这些变体中,只需要保留其中之一即可,其它变体的单元将用这个变体来替换。删除绝对冗余变体基本不会降低合成效果。

绝对冗余,必须满足如下的严格限制:

(1) 韵律的限制(Prosody Restrict). 在单元不定长树的叶子节点下,韵律按重要性从高到低,限制是非常严格的,因此绝对冗余的判定一般在叶子层次进行。

(2) 声学限制(Acoustic Restrict). 在声学表现上,为了能够确保单元的互换性,本文将变体的 LSF 和 Pitch 的距离严格限制在一定的范围以内。即两个变体的声学距离(由 2.4 节定义)必须在同一个小的邻域内:

$$DIS(u_1, u_2) < \alpha,$$

这样的变体才被认为是满足条件 2。

(3) 规则限制(Rules Restrict). 对于单元不定长树,其仍可能存在一定的不合理性。所以需要在判定的过程中,根据先验知识建立一套规则库严格约束判定条件。如替换句头的必须也是句头的;如果后接浊音,用来替换它的变体的也必须是后接浊音等等。

这些可以替换的变体之间,有的是两两可以互换,也有的是某两个单元能够被第三个单元替换但是这两个单元却不能互换。诸如此类的替换关系形成了一个复杂的网状替换关系图。本文适当简化这个图关系,利用 2.4 节的距离度量进行聚类,选取类中心作为最佳的变体保留,得到替换关系表。绝对冗余的裁剪被本文称为 ARE(Absolute Redundancy Elimination)。

定义 1. 同时满足 Prosody Restrict, Acoustic Restrict 和 Rules Restrict 的变体,经过聚类除去聚类中心的那些变体即被称为绝对冗余。

某个单元的变体,KBCE 合成系统最多预选(pre-select) K 个,如果该单元变体数目大于 K ,总有一些变体不被使用,除去这最具有代表性的 K 个单元。其它的就称为相对冗余。

相对冗余可以被删除。但是因为每次合成的时候预选的 K 个变体可能不相同,而删除相对冗余,导致可以被预选的不定长变体数目减少,即预选的范围变小了,合成效果将可能有一定损失。这要求必须非常小心的处理相对冗余的删除。本文采用如下方法,在多个变体中选择最具有代表性变体,而删除相对冗余:

对于单元索引树的某个结点,若 $K \leq \text{变体数} \leq G_b$,使用聚类方法来选择其中变体,也就是将变体聚为固定的类数,保留每个类中心来代表该类的所有变体。这里距离度量参见 2.4 节。 G_b 的设置是为了变体聚类更为精细一些,不至于在太大范围内进行聚类,从而使得变体的韵律和声学环境的损失被有效控制。本文称相对冗余的裁剪为 RRE(Relative Redundancy Elimination)。

定义 2. 对于单元索引树的某个节点,若其所含变体数量满足: $K \leq \text{变体数} < G_b$ 且节点的深度大于最小阈值 L ,除去聚类中心之外其它变体被称为相对冗余。

在实际声学聚类时候,发现有一些孤立变体离各个类中心距离都很大,其中一部分经过测听为发音不好的变体,这些变体不仅对合成没有帮助,反而还有损害,因此必须将这些坏变体删除。本文利用文

献[3]的结论,对这些孤立变体进行自动声学判定,找出那些坏变体(Auto Bad-Variants Elimination, ABVE)并且删除。

定义 3. 在寻找相对冗余中的聚类孤立点中,进行基于听感量化自动声学判定后,得到的变体为坏变体。

由本小节分析可知,冗余变体的裁剪就是要删除满足定义 1~3 的三种变体。

2.3 裁剪的递阶处理

语音库与一般数据不同之处在于:如果把音节数看成不定长的阶(粒度),那么语音库中的变体按照这种阶(粒度)构成了一种层次化的逻辑结构。举例来说,如不定长变体 $V_1V_2V_3$ 包含了不定长变体 V_1V_2 , V_1V_2 又包含了变体 V_1 。如果聚类只是对单音节进行,那么 V_1V_2 的关联性、 $V_1V_2V_3$ 的关联性信息都会丢失,即高阶不定长信息会被破坏。不定长方法是基于语料库的语音合成的一个关键技术,如果裁剪处理得不好,重要的不定长损失太多,合成的效果就会严重下降。语音库裁剪必须致力于避免重要不定长的损失。本文采用了按照变体的不同粒度进行递阶聚类的方法(这里的递阶聚类和文献[17]中的层次聚类不同),来减小重要不定长的损失:

首先对高阶不定长采用 2.4 节的距离进行聚类,并将聚类中心投影(Project)到相关联的低阶变体上。然后在对低阶变体聚类时,让当前聚类中心更倾向于高阶类中心的低阶投影。这里的所谓“倾向”可采用两种方式:第一种方式,所有低阶变体都有一个权重 w_i ,且 $\sum w_i = 1$, $w_i > 1$ 。投影变体的权重较大,然后每次以 w_i 为各变体的出现概率随机产生变体,这样重复足够多次,然后进行样本聚类;第二种方式,直接优先保留投影变体,不足的再进行聚类取中心。我们采用了这两种方式进行了实验,结果差别不大。基于实际计算代价的考虑,第二种方式虽然简单但却同样有效。图 2 给出了变体递阶聚类的一个

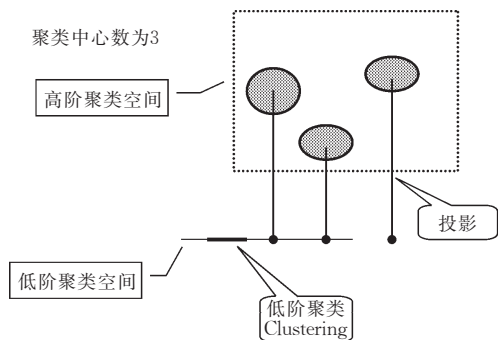


图 2 递阶聚类的图示

示意。

递阶聚类的方法,使得重要的高阶不定长(聚类中心)在低阶聚类时,也被作为中心保留了下来。这一方法是有效减小不定长损失的关键所在。第 3 节的测听实验也证明了这一点。

2.4 聚类的距离度量

如何度量变体间距离呢?这同样也是一个关键的问题,它直接决定着裁剪后合成的效果。最好的方式是直接以合成效果为依据来指导裁剪,也就是找到直接反映听感的声学依据,并以这种声学依据作为某种距离度量。

在听感量化的研究^[3]中我们发现,对于变体 u_1 和 u_2 ,如下的距离定义可以比较好的反映它们在听感上的差异:

$$DIS(u_1, u_2) = \sqrt{DIS^2(x_1, x_2) + w_1 \times DIS^2(y_1, y_2)},$$

其中以 x_1 和 x_2 描述变体 u_1 和 u_2 的超音段特征,而 y_1 和 y_2 描述变体 u_1 和 u_2 的音段特征。

两个声学变体的超音段特征对听感造成的差异,可以近似用它们 F_0 参数(变体的每个数据帧的 F_0 值形成矢量,作为该变体的 F_0 参数)的均值和包络的欧式距离来度量:

$$DIS(x_1, x_2) = |mean(x_1) - mean(x_2)| + EuclidDist(x'_1, x'_2),$$

其中, $EuclidDist(x'_1, x'_2)$ 是指 x_1 和 x_2 去均值后的数据之间的欧式距离。

自然度的影响不仅仅是由于超音段特征引起的,音段上的差异同样会造成自然度的下降。音段特性是通过线谱对参数(LSF 参数)进行描述的,LSF 参数是线性预测(LPC)参数的一种表现形式,线性预测是一种源-滤波器方法^[18-19],可以用来分析和合成语音信号。这种方法假定信号当前样点值可以被以前 P 个样点值通过线性组合得到,其数学表达式如下式:

$$s(n) = \sum_{k=1}^P a_k s(n-k) + Gu(n) \quad (1)$$

在上式中, P 是分析的阶数, a_k 是线性预测系数(LPC 系数), $u(n)$ 是激励信号, G 是相应的增益;在合成语音时,对于浊音信号,可以用一段周期性脉冲信号作为相应的激励,对于清音信号,以一段白噪声作为激励信号。式(1)相应的 z 变换形式如下式:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^P a_k z^{-k}} \quad (2)$$

其中, $S(z)$ 和 $U(z)$ 分别代表语音信号和激励信号

的 z -变换形式, $H(z)$ 表示声道相应滤波器;

LSF 系数是由 Itakura 提出的一种 LPC 系数的另一种表达形式, 假定 P 阶的 LPC 滤波器由式(2)表达, 可以根据式(2)中的 P 阶 LPC 滤波器 $A^P(z)$ 构造一个 $P+1$ 阶多项式 $P(z)$ 和 $Q(z)$ 如下^{[20]①}:

$$P(z) = A^P(z) + z^{-(P+1)} A^P(z^{-1}) \quad (3)$$

$$Q(z) = A^P(z) - z^{-(P+1)} A^P(z^{-1}) \quad (4)$$

其中, $P(z)$ 和 $Q(z)$ 分别指奇数阶和偶数阶 LSF 参数.

LSF 系数和共振峰具有很好的吻合性: 奇数级 LSF 系数和共振峰表现出很强的映射; 而奇数级和偶数 LSF 系数之间的距离和谱的谐波特性吻合较好, 当奇数级和偶数 LSF 系数较小时, 频谱上表现出强烈的谐波特性.

听感量化的研究表明, 可以通过计算两个单元的 LSF 参数之间的欧式距离, 来度量两个声学变体在音段上的差异, 和 F_0 的距离度量方法不同, 两个变体之间的 LSF 参数 y_1 和 y_2 之间距离计算方法要复杂些, 文献[14]有对两个变体 LSF 参数之间距离计算方法的详细介绍, 为本文所采用.

同时引入权值 w_1 , 用来调节超音段特征距离和音段特征距离在最后声学距离中所占的权重, 使距离计算结果和主观测听的结果对应得更好. 实验表明 w_1 为 0.6 比较好.

基于听感量化的距离度量, 能较好地反映听感与声学之间的对应关系: 若变体 u_1 和 u_2 的听感差异较大, 它们的 $DIS(u_1, u_2)$ 值也较大; 若变体 u_1 和 u_2 的听感差异较小, 它们的 $DIS(u_1, u_2)$ 值也较小. 以这种距离度量进行聚类, 实际上就是间接以听感作为依据进行聚类, 从而保证了裁剪后合成效果不会严重下降.

2.5 NuClustering-VPA 算法的描述

根据以上的基本思想, 本节将介绍语音库裁剪的不定长递阶聚类算法 NuClustering-VPA (Non-uniform-Clustering based Variant Pruning Algorithm). NuClustering-VPA 的形式化描述如下:

1. 对语音库进行坏单元剔除, 即执行 ABVE 并进行绝对冗余替换 ARE:

$Instances_Set = ABVE(Instances_Set);$

For all Unit_trees do:

① $Instances_Set = Prosody_Restrict(Instances_Set)$

② $Instances_Set = Accoustic_Restrict(Instances_Set)$

③ $Instances_Set = Rules_Restrict(Instances_Set)$

④ $Instances_Set = ARE(Instances_Set)$

2. 裁剪掉所有声学变体数少于 G_a 的语音单元(URE);

$Units_Set = URE(Units_Set);$

3. 进行相对冗余的裁剪(RRE):

For $L = Max_Length$ to 1 $\forall Unit, Length(Unit) = L,$
 $NuClustering(Unit.RootNode)$

即 L 从不定长的最大长度开始到 1 (为单音节的长度):

对所有长度为 L 的单元 $Unit$, 执行

$NuClustering(Unit.RootNode),$

此处 $Unit.RootNode$ 为单元树 $Unit$ 的根节点.

NuClustering-VPA 算法第一步中, ABVE 和 ARE 所用到的聚类方法和 RRE 中的聚类方法是一样的. NuClustering-VPA 最重要的过程是 RRE. 而 RRE 中最重要的子函数 NuClustering 和 Tailoring 如图 3 和图 4 所示.

```

NuClustering(Node)
{
  if Node is NOT 叶子
  {
    if Level < 最小层数 L
    {
      Node = Node 的孩子
      NuClustering(Node)
    }
    else if Node 所含变体数 > Gb
    {
      Node = Node 的孩子
      NuClustering(Node)
    }
    else if Node 所含变体数目 > K
    {
      Num = 保留率 × Node 所含变体数
      Tailoring(Node, Num)
      return
    }
    else return
  }
  else if Node 所含变体数目 > K
  {
    Num = 保留率 × Node 所含变体数
    Tailoring(Node, Num)
    return
  }
  else return
}

Tailoring(Node, Num)
{
  去除高阶中心在当前的投影变体
  Clustering(剩余变体, Num-投影变体数)
  保留聚类中心, 删除其它变体
}

```

图 3 NuClustering 和 Tailoring 函数的形式化描述

从图 3 和图 4 中可以看出 NuClustering 的聚类要求:

(1) 在单元树的一定深度上进行 ($Level > L$);

(2) 参与聚类的节点数必须达到一定要求 ($K < Node$ 所含节点数 $\leq Gb$);

这样就减小了单元的声韵环境因为裁剪而发生损失的可能性.

Tailoring 函数每次进行聚类时, 都会首先保留高阶中心在当前的投影, 然后再对剩余变体进行聚类, 这样高阶不定长变体聚类的结果, 就会直接被

① ITU-T, INTERNATIONAL TELECOMMUNICATION UNION, 1995

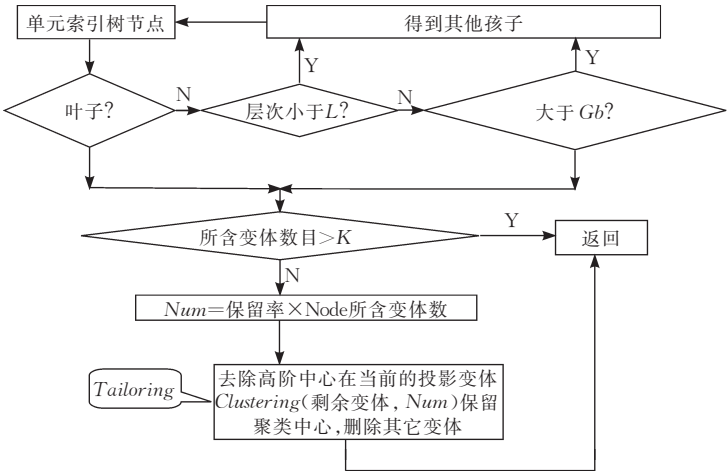


图 4 NuClustering 的流程图

保留在低阶不定长变体的聚类结果中,因此有效地减小了裁剪后不定长的损失. 其中聚类可以使用 LBG 推挤方法^[15-17]或者 K-均值聚类方法^[15-17].

从整体上讲,NuClustering-VPA 的聚类从高阶不定长开始,然后粒度逐渐变小到低阶不定长;最后所有长度的不定长,都会直接反映在单音节聚类上. 这种方法一定程度保存了各种粒度上重要的不定长(各阶的聚类中心). 要比直接在单音节单一层次上聚类更好,第 3 节的测听结果也显示出了这一点.

3 测听的结果和分析

本文对 KBCE 使用的一种原始库(语料库来源于数年的人民日报)进行不同规模的裁剪,然后采用覆盖率最大的中文语料 150(包括开放性和限定性)句语料,对合成效果进行了大规模的主观测听. 主观度量采用目前通行的 5 分制 MOS(Mean Opinion Score)^①分.

本文使用的 150 句语料是通过文献[3]的语料搜索方法得到(文献[3]的语料搜索方法寻找的是覆盖率最大的语料,也就是最能从语言学、韵律学、声学方面代表中文语言现象的句子. 根据我们的实验研究经验,同时结合测听员疲劳度考虑,使用这种方法在限定性预料中得到 50 句、在开放性语料中得到 100 句进行测听具有较好的代表性). 前 100 句为在开放性语料(来源不限于《人民日报》,包括各种类型的语料)中覆盖率最高的新文本,后 50 句为在限定性语料(来源只限于新的《人民日报》语料)覆盖率较低的新文本. 由 5 位测听员测听(测听员事先不知道系统的对应关系,每次各系统相同的句子出现的顺序是随机的),给出 MOS 分.

表 1 给出了各种裁剪率的系统的单音变体和高阶(长度,即音节数大于 1)不定长的保留情况. 这里裁剪率=1-保留率,保留的所有低阶(单音)和高阶不定长都是各阶的聚类中心,是各个语音单元最重要的声学变体.

表 1 各裁减规模的系统

系统	保留单音变体数	被保留的高阶不定长数	保留率/%
A	509477	322989	100
B	412736	180401	81.01
C	374000	150793	73.41
D	318000	112590	62.42
E	201926	37638	39.63

表 2、表 3 分别给出了 100 句和 50 句测听的 MOS 分. 可以看出裁剪率较高时 MOS 下降得并不多,即使在语音库只保留 39.63%时,MOS 降低也没有超过 0.1. 说明裁剪处理中的递阶方法和聚类的距离度量是十分有效的. 图 5 为 MOS 分随保留率的变化曲线,图 5 十分清楚地显示出了 MOS 分随保留率的变化.

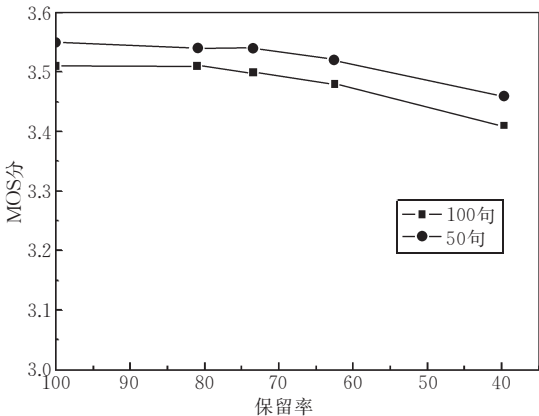


图 5 MOS 分随音库大小的变化

① ITU-T Recommendations P. 10, P. 800, and P. 800. 1, (<http://www.itu.int/rec/recommendation.asp?type=items&lang=e&parent=T-REC-P.800-199608-1>)

变化幅度很小. 这个结果表明, NuClustering-VPA 不仅比一般音节或音素的聚类裁剪方法的 MOS 分曲线下下降要慢, 相比于其它裁剪方法在保留率在 50% 以上才不会有严重下降, NuClustering-VPA 即使在保留率为 39.63% 时, MOS 分下降也没有超过 0.1.

表 2 前 100 句的 MOS 分

系统	测听员 1	测听员 2	测听员 3	测听员 4	测听员 5	MOS
A	3.41	3.45	3.8	3.55	3.34	3.51
B	3.41	3.46	3.845	3.53	3.28	3.51
C	3.405	3.485	3.815	3.52	3.29	3.50
D	3.37	3.425	3.795	3.53	3.28	3.48
E	3.32	3.44	3.785	3.295	3.235	3.41

表 3 后 50 句的 MOS 分

系统	测听员 1	测听员 2	测听员 3	测听员 4	测听员 5	MOS
A	3.34	3.68	3.87	3.48	3.37	3.55
B	3.42	3.71	3.84	3.42	3.31	3.54
C	3.35	3.68	3.84	3.49	3.34	3.54
D	3.38	3.68	3.79	3.41	3.35	3.52
E	3.34	3.63	3.79	3.24	3.31	3.46

从实验结果可以看出, NuClustering-VPA 即使将语音库裁剪到 39.63% 时, 合成的自然度几乎没有发生显著下降, MOS 降低没有超过 0.1. 本文是在 KBCE 这样一个具有高自然度语音合成系统上进行裁剪(请参见脚注②), 其中使用了大量的不定长变体, 如果裁剪不得当, 即使只有少数重要不定长损失, 合成的自然度也会有比较明显的下降^[3]. 因此可以看出, NuClustering-VPA 的裁剪结果有显著的意义: 它保留了对合成最为重要的高质量不定长变体, 而裁剪了冗余和次要的不定长变体.

4 结论和下一步工作

本文提出 NuClustering-VPA 算法: 对不同粒度的不定长变体进行递归聚类, 根据高阶聚类结果调整低阶变体的聚类, 使得低阶聚类中心有所偏向. NuClustering-VPA 算法保留那些对于不同粒度不定长来说, 在声韵上最为重要的变体, 从而减小了裁剪对不定长的破坏. NuClustering-VPA 算法在 KBCE 系统上得到了实现. 测听实验表明, 在 KBCE 这样的高自然度合成系统上, 即使语音库裁减率为 39.63% 时, 合成自然度下降仍然较小, 并保持在较高的水平. 科大讯飞公司利用这一技术, 推出了面向桌面的语音系统——文语通, 这一系统得到了用户的肯定.

这里我们提出下一步工作: 首先 NuClustering-

VPA 算法依赖于语音距离的度量. 实际上, 完全可以利用合成中 Vertibi 挑选算法来衡量变体的重要性, 使得裁减和合成方法紧密结合起来, 而不依赖于具体的距离度量. 这样做的另一个好处是: 当变体挑选方法发生改变时, 裁剪也可以作相应的变化, 使得裁剪更灵活. 其次, 聚类算法保留了最为重要的不定长, 但是对损失掉的不定长没有作处理. 实际上, 损失的变体可通过存在的变体进行一定的替换和虚拟, 即所谓虚拟不定长. 这些都是值得进一步研究的, 也是我们下一步的工作.

参 考 文 献

[1] Hunt A, Black A. Unit selection in a concatenative speech synthesis system using a large speech database//Proceedings of the ICASSP1996. 1996, 1: 373-376

[2] Sagisaka Y, Kaiki N, Iwahashi N, Mimura K. ATR-v-TALK speech synthesis system//Proceedings of the ICSLP 1992. 1992, 1: 483-486

[3] Liu Qing-Feng. Speech synthesis study based on perception quantification[Ph. D. dissertation]. University of Science and Technology of China, Hefei, 2003(in Chinese)
(刘庆峰. 基于听感量化的语音合成研究[博士学位论文]. 中国科学技术大学, 合肥, 2003)

[4] Chu M, Peng H, Yang H, Chang E. Selection non-uniform units from a very large corpus for concatenative speech synthesizer//Proceedings of the ICASSP2001. 2001

[5] Rabiner L R. A tutorial on hidden markov models and selected application in speech recognition. IEEE Proceedings, 1989, 77 (2): 257-285

[6] Breiman L, Friedman J, Olsen R, Stone C. Classification and Regression Trees. Pacific Grove, CA: Wadsworth & Brooks, 1984

[7] Black A W, Taylor P A. Automatically clustering similar units for units selection in speech synthesis//Proceedings of the Eurospeech1997. 1997, 2: 601-604

[8] Hon H, Acero A, Huang X, Liu J, Plumpe M. Automatic generation of synthesis units for trainable text-to-speech systems//Proceedings of the ICASSP1998. 1998, 1: 293-296

[9] Kim S H, Lee Y L, Hirose K. Pruning of redundant synthesis instances based on weight vector quantization//Proceedings of the Eurospeech2001. 2001: 2231-2234

[10] Kim S H, Lee Y L, 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//Proceedings of the ICSLP2002. Denver, 2002: 125-128

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

[13]

Ling Z H, Hu Y, Shuang Z W, Wang R H. Compression of speech database by feature separation and pattern clustering using STRAIGHT//Proceeding of the ICSLP 2004. 2004; 766-769

[14]

Wouters J, Macon M. Control of spectral dynamics in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing, 2001, 9(1): 30-38

[15]

Shi Zhong-Zhi. Knowledge Discovery. Beijing: Tsinghua University Press, 2002(in Chinese)
(史忠植. 知识发现. 北京:清华大学出版社, 2002)

[16]

Pieter Adriaans, Dolf Zantinge. Data Mining. England: Addison-Wesley,1996

[17]

Han Jia-Wei, Kamber Micheline. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000

[18]

Yang Xing-Jun, Chi Hui-Sheng. Digital Speech Signal Processing. Beijing: Publishing House of Electronics Industry, 1995(in Chinese)
(杨行峻,迟惠生. 语音信号数字处理. 北京:电子工业出版社,1995)

[19]

Chen Yong-Bin, Wang Ren-Hua. Speech Signal Processing. Hefei: University of Science and Technology of China Press, 1990(in Chinese)
(陈永彬、王仁华. 语音信号处理. 合肥:中国科学技术大学出版社,1990)

[20]

Esther Klabbers. Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing, 2001, 9(1): 39-51



ZHANG Wei, born in 1975, Ph. D. , associate professor. His main research interests include machine learning and statistical analysis, scalable TTS system.

WU Xiao-Ru, born in 1972, Ph. D. . His research interests include speech recognition and TTS.

LIU Jiang, born in 1980. His research interests include speech recognition and TTS.

WANG Ren-Hua, born in 1943, professor, Ph. D. supervisor. His research interests include speech recognition and TTS, signal processing.

Background

This work is from the project of Scalable Speech Synthesis (Text to Speech) System, which is supported by the National Natural Science Foundation of China (60602017) and the National High Technology Research and Development Program (863 Program) of China (2004AA114030). The research aims at the theories and key techniques of pruning corpus redundance and of making the TTS system scalable to hardwares. The iflytek research team who carry out this pro-

ject has constructed a Corpus-Based Continuous Chinese-English Text-to-Speech Engine, which is awarded best performance in the routine national “863” evaluation.

This paper solves the question of Tailoring TTS font or pruning redundant synthesis instances without severely degrading naturalness scored by MOS. Thus the result can be used to analysis redundance of large corpus and to shrink database of synthesis instances according to hardwares.