中 文 信 息 学 报
第 18 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 3

文章编号:1003 - 0077(2004)03 - 0054 - 07

嵌入式汉语 TTS 系统的设计与实现*

刘 涛,叶振兴,蔡莲红

(清华大学 计算机科学与技术系,北京 100084)

摘要:针对手持设备和 PDA 存储量较小的特点,本文提出了基于音节基频包络特征、采用 k 中心点算法聚类裁减音库容量的方法。聚类结果的听辩实验和统计分析表明此算法可以保证聚类内部音节样本的相似性及类间样本的相异性。经过对汉语可选合成基元的分析,系统中首次引入声韵母半音节与音节作为混合基元,构造了基于混合基元的音库。经过对样本集分别聚类裁减,进一步压缩了音库容量,并在 PDA 平台上实现了嵌入式 TTS 系统。

关键词:计算机应用;中文信息处理;嵌入式;汉语语音合成系统;音节聚类;混合基元

中图分类号:TP391 **文献标识码**:A

Design and Implementation of Embedded Chinese TTS System

LIU Tao, YE Zhen-xing, CAI Lian-hong

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

Abstract: Aiming at handset devices with small memory, we employ k-medoids algorithm with pitch contour as the feature to reduce the size of the speech corpus of the current Chinese TTS system. The result of objective evaluation and statistic analysis shows that the similarities of the samples in a same cluster and the dissimilarities in different clusters can be guaranteed. In this system, hybrid units composed of the semi-syllable units of initial and final and the conventional syllable units are used to construct the new corpus according to the analysis of the probable units in Mandarin TTS system. After the sample sets of the hybrid units are reduced by clustering algorithm respectively, the embedded Chinese TTS system is implemented on the PDA platform.

Key words: computer application; Chinese information processing; embedded TTS system; corpus reducing; hybrid synthesis unit

1 引言

近些年语音合成技术取得了突破性的进展,汉语语音合成的可懂度与自然度有了较大提高,目前已经开始走出实验室,进入市场实用阶段,其中一个重要的发展趋势就是把语音合成系统小型化,使其可以运行于 PDA、手机等移动设备。在这些移动设备上,屏幕面积小,阅读不方便,资源有限,形成了信息获取的瓶颈。因此,如果我们能够在 PDA 等设备上应用语音合成技术,通过"听"而不是"看"来获取信息,就可以打破瓶颈的束缚,为移动设备提供更加自由方便的交互界面。同样在信息家电等领域嵌入式 TTS 系统也将有其用武之地。

TTS 技术结合了自然语言理解、语言学、语音学、声学、计算机科学等多个领域的研究成

* **收稿日期**:2003 - 05 - 26
基金项目:国家自然科学基金资助项目(60275014);863 资助项目(2002AA117010 - 05)
作者简介:刘涛(1977 —),男,硕士,主要研究方向为嵌入式语音合成.

果。目前汉语 TTS 系统都采用基于大规模语音语料库的波形拼接的方法[1]。其中音库一般在几百 MB 到几个 GB 之间,所以实现嵌入式 TTS 系统的关键在于音库容量。摩托罗拉公司的陈芳等人以基频为特征、采用分段变长量化的方法裁剪音库容量[2]。中科院讯飞实验室的双志伟等人建立基于韵律词的量化韵律模版库,依赖韵律调整实现小型化合成系统[3]。中科院声学所的孙金城等人以基频作为特征采用 k 均值聚类算法裁减音库[4]。我们在现有大型 TTS 系统的基础上重新设计并实现嵌入式 TTS 系统,其遵循以下原则:

1) 降低现有 TTS 系统的空间复杂度。其主要任务是裁减目前系统音库容量,使之适应嵌入式系统需求。

2) 降低现有 TTS 系统的时间复杂度。目前 TTS 系统中采用的文本分析、基元选取、语音合成的算法都比较复杂,需要进行优化或者选择复杂度较低的算法代替。

3) 保证一定的语音合成质量。在降低现有 TTS 系统时间和空间复杂度的基础上,不能过多损失合成质量。在保证可懂度的基础上,也要保证合成语音比较自然。

本文将首先给出以音节基频包络为特征的音节样本集聚类算法,通过此算法可以有效的裁减音库容量;在此基础上,本文又引入声韵母半音节与音节共存作为混合基元,实现了小容量高质量的嵌入式 TTS 系统。

2 基于音节基频包络特征的样本集聚类

目前的汉语 TTS 系统采用有调音节作为合成基元,在音库中每个基元有数目不等的样本,一般来说,使用频率较高的音节具有较多的样本。通过对音库的听辩分析,可以发现音库中很多样本在听感上非常相似,也就是说在合成语音时,这些样本可以互相替换,那么我们就可以对其进行裁减。另一方面,为了保证合成质量,保留的样本也应该代表不同的语境,这样才能保证基元选取算法可以选到合适的样本。本文将以音节的基频包络为特征、采用 k 中心点算法对音节样本集进行聚类压缩,从而合理的裁减音库容量。

在聚类特征选取上,我们研究了 MFCC、基频、时长等参数。通过听辩实验,分析比较了不同参数的聚类效果。最后我们选取基频包络作为聚类特征,并且基于时长参数进行了归一化处理。采用基频包络为特征的聚类方法简便易行,聚类结果表明不同类别可较好地反映听感上的差异,基本上满足了我们的要求。

2.1 聚类算法

本文使用最为广泛的基于划分的方法对音节样本集进行聚类,其中最为著名也最为常用的划分方法是 k- 均值与 k- 中心点算法[5]。由于 k- 均值算法对于孤立点是敏感的,一个有极大值的样本可能会相当程度扭曲聚类的分布,所以本文采用了基于代表性样本的 k- 中心点算法。k- 中心点算法采用类最靠近中心的样本而非计算出的均值作为新的类中心,它仍然基于最小化所有样本与中心样本之间的相异度之和的原则来执行。其基本策略是:首先为每个类随意选择一个代表点即类中心;剩余的样本根据其与代表样本的距离分配给最近的一个类。然后反复的用非代表样本代替代表样本,以改进聚类的质量。它具有如下特点:1. 能有效的处理中等规模的数据集;2. 对孤立点不敏感;3. 算法的执行结果和样本的顺序、初始代表点的选择有关;4. 计算复杂度较高为 $O(k(n-k)^2)$,其中 k 为聚类数目,n 为样本数目。

2.2 聚类实验及结果分析

为了验证此算法是否符合要求,我们设计了一个聚类实验,并且通过主观听辩实验和聚类结果的声学、韵律学分析来验证聚类效果。实验中选用所有去声音节作为实验对象,并且首先

分为清声母、零声母以及浊声母去声音节三大类分别处理。图 1、2 分别为其中清声母、零声母去声音节的聚类结果类中心基频曲线图。经过对聚类结果的统计分析和主观听辩实验结果分析,我们得到以下结论:同类内样本听感近似,具有互换性,而类间差异明显;同类内样本在语句中位置以及韵律结构边界的分布上具有相对一致性。由于聚类只裁减在听感上近似的样本,从而保证了裁减后的音库在声学特性上损失较小;另一方面,聚类对应的语境同样具有代表性,从而保证了压缩后的音库在语境上比较完备。因此,我们可以利用基于基频包络的音节聚类来对目前的大语料库进行裁减。

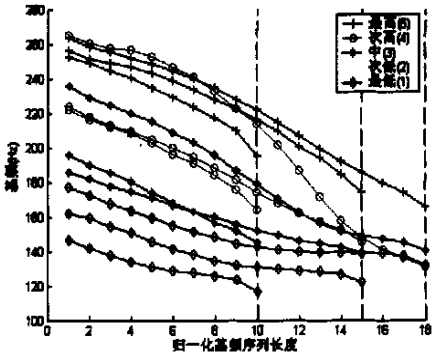

图 1　清声母去声音节聚类结果

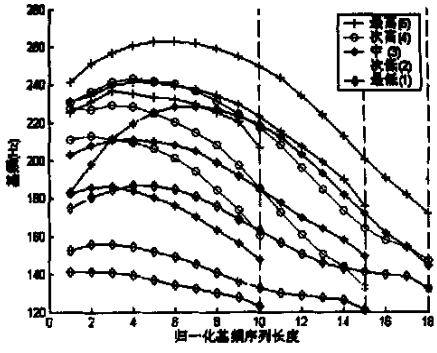

图 2　零声母去声音节聚类结果

3　基于混合基元的嵌入式 TTS 系统设计

基于基频包络的音节聚类可以很好的裁减 TTS 系统音库的容量,但是其压缩能力有限,当音库容量下降到一定程度时,由于样本数量过少,其合成质量也会明显下降。因此本文采用重新设计合成基元的方法进一步裁减音库容量。

3.1　基元集设计

我们要在现有的大音库基础上裁减音库容量,主要的方法有:

® 减少基元类别数。目前音库采用有调音节作为基元,其类别数目在 1300 左右,而每种基元又需一定数量的样本才能保证合成质量。若减少基元数目,如采用无调音节,音库容量随之下降。

® 减少基元样本数。上述的样本集聚类是其方法之一。

® 缩小基元单位。汉语语音合成系统中除音节外可以选择的语音学单元包括:半音节(Semi-Syllable)、声韵母(Initial/Final)、音素(Phone)等。

音素在英语合成系统中得到了广泛的应用,并取得了非常好的合成效果。音素给标注与声学描述带来了一定的困难。另一方面基于音素的合成不可避免的要借助于韵律修改,但是目前汉语的韵律修改模块还不成熟,容易产生噪音。

声韵母是汉语特有的一种发音结构。一般来说一个声母和一个韵母组成音节,也有零声母音节。声韵母需要从音节中切分得到,但是很多时候这种分界点并不明显。考虑到韵母的发音比较稳定,所以把声韵过渡部分归入声母,也就形成了声母 + 韵头的半音节,简称声母半音节,而后面的韵母段则称为韵母半音节。

声韵母半音节基元是适合汉语特点的一种合成基元,使用这种基元具有以下优点:

® 使用声韵母半音节作为合成基元,上下文相关信息也变得简单而确定。比如,与声母相接的只能是韵母或者静音,而与韵母相接的也只能是声母或静音或韵母,而且,韵母左边相

接的声母只能是与其搭配起来能够形成汉语音节的那些声母。所以,上下文相关的声韵母数目也远远少于目前的上下文相关的音节数目。

® 声韵母是汉语音节所特有的一种结构,有很多关于声韵母的语音学方面的知识和研究成果可以被我们采用。

® 选择声韵母作为基元,它的语音段长度,以及基元数目都是比较适当的。如果不考虑上下文信息,声韵母一共只有 59 个,其中声母 21 个,韵母 38 个。

因此,我们将引入声韵母半音节作为新的合成基元构造嵌入式 TTS 系统。

3.2 混合基元音库的构造

声韵母半音节切分自音节,但是部分音节的声韵母半音节切分是十分困难的,这与声母的类型有关。其中浊声母 l, m, n, r 和塞音不送气声母 b, d, g 开头的音节通常不易切分。如果切分点不准确,那么拼接合成的音节就不自然,这会在很大程度上影响合成语音的质量。对于这些音节,我们不进行切分,而保留整体音节作为基元,零声母音节也将同样处理。所以,在我们的新音库中,音节将与声韵母半音节共存作为混合基元。

1) 声韵母半音节的收集

基于混合基元的音库将在目前基于音节的音库的基础上构造。这主要因为:

® 如果重新设计音库需要经过文本设计、音频数据录制、切分标注等等工作,耗费大量的人力、物力和时间。

® 基于音节的音库具有良好的设计和比较精确的语音学、语言学标注。这为我们快速、准确的建立新音库提供了很好的基础。

如何确定切分点是声韵母半音节切分需要处理的关键问题。这里我们利用了音库中关于音节韵母部分基因周期的标注信息解决。目前的音节标注信息给出了音频数据中周期信号部分的每个峰值点的采样点序号。一般清声母开头的音节,如音节"sha1"的声韵具有明显的分界点,声母部分不具周期性,而韵母部分则呈周期性变化,其峰值点标注全部处于韵母部分,第一条标注线前的零点基本可以认为是声韵母半音节的切分点,且其中的过渡部分归声母半音节所有。但音节"ma1"的声韵分界并不明显,其声母部分信号也呈现周期性变化,这是因为汉语具有声母浊化现象,同时其峰值点标注也容易出现不准确的情况,这样就不能利用基频标注信息进行声韵切分。

对于清声母开头的音节,我们可以利用其基频标注信息找到切分点,完成音节的声韵切分,得到声韵母半音节基元;而对于其它音节则不进行切分处理,也就形成了音节基元。基元的具体构成如表 1 所示。

表 1 基元集构成

基元	基元集构成
声母半音节	c, ch, f, h, j, k, p, q, s, sh, t, x, z, zh
韵母半音节	所有可与声母半音节搭配成音节的韵母
音节	以 b, d, g, l, m, n, r 为声母的音节;零声母音节

2) 混合基元样本集的裁减

对基于混合基元的音库样本集的裁减将按照不同的基元类型进行,其基本方法还是基于音节或半音节的声学特征对样本集进行聚类,从而减少样本数量,达到裁减音库容量的目的。

® 音节样本集裁减

即采用以基频包络为特征的样本集聚类算法进行裁减。

® 声母半音节样本集裁减

音节的声母段一般不体现声调信息,对于清声母尤其明显,因此我们忽略声调,把拼音相

同的音节的声母作为一个集合进行聚类。我们直接利用声母的采样点就行聚类。聚类操作分为两步,与音节样本集的聚类相似,首先按照声母时长分类,然后在每个时长分类中进行时长归一化再采用 k 中心点算法聚为指定的类数。

® 韵母半音节样本集裁减

根据韵母所处原音节的声母发音方法做了进一步分类,也就是把原音节声母类型相同且声调相同的一类韵母半音节作为一个集合,在这个集合内部进行聚类,减少样本数量,裁减音库容量。经过听取不同声韵的搭配实验以及声母的发音方式,我们对声母的分类如表 2 所示。

表 2 声母分类表

声母分类	z, c, s	zh, ch, sh	j, q, x	k, h	p, t, f

因为聚类的范围扩大到韵母半音节,所以就有可能进一步压缩音库的容量。

3.3 基于混合基元的基元选取与拼接合成

由于现有 TTS 系统的各个模块都经过了精心设计、实现与长时间的实践检验,已经相当成熟,因此我们尽量利用现有 TTS 系统的相关模块。实现中先由声韵母半音节拼接成音节,再由音节拼接成句子的合成策略。

目前的汉语 TTS 系统一般都采用基于语境信息的基元选取策略。我们现有的 TTS 系统就包含了对音节详细的语境信息标注:音段参数、位置参数和关联参数,在进行基元选取时,采用了全匹配代价模型的韵律代价函数[6]。

为了利用现有的基元选取算法,我们保留了原有音库中的所有音节标注信息;在基元选取时,依然按照音节的语境信息和原有算法进行选取;对于选中的音节如果在新音库中是以音节的形式存在,则直接选取此样本所在聚类的保留样本进行波形拼接,否则就从库中选取相应的声韵母半音节(分别为声母与韵母半音节所在聚类的保留样本)拼接得到音节,再进行句子的波形拼接。

这里的拼接合成处理是指声韵母半音节拼接合成音节模块,而音节拼接成句子的处理将继承现有 TTS 系统中的波形拼接模块。由于在声韵母切分时我们把过渡段归入声母部分,保证了在聚类结果中各个聚类中心包含了比较丰富的上下文信息,这样声韵母拼接时不需要做太多的韵律修改。出于对 TTS 系统合成的实时性需求的考虑,我们采用了复杂度较低的加窗平滑拼接算法。实验结果表明合成结果的质量没有明显的下降。

4 嵌入式 TTS 系统的实现

4.1 目标平台

嵌入式 TTS 系统主要应用于移动办公及生活领域,如汽车上的语音导航系统,手机上的语音提示功能等。目前普遍应用的嵌入式硬件平台主要是 PDA 和手机,其中 PDA 配置较高,计算能力更强,内存配置更多,且可以运行目前主流的嵌入式操作系统,是理想的嵌入式 TTS 系统的实现平台。我们最终选取 Compaq 公司的 iPaq 3660 作为目标平台,它配置了 206MHz 的 CPU 和 64M 内存,运行 WIN CE3.0 操作系统。由于嵌入式系统的快速发展,目前已经出现了更多更高配置的系统平台。

4.2 样本数分配和保留策略

在我们的 TTS 系统中,统计了每个基元样本被选中的频度。经过对大约 100MB 的文本语料作合成统计,我们得到了音节样本集中每个样本的使用频度以及每个音节的出现频度,这为我们裁减音库音节样本集提供了两个重要的参数。

首先,无论是对音节样本集的聚类还是对声韵母样本集的聚类来说,最终都需要选取一个样本保留下来代表此聚类。一般情况下选取聚类中心样本,也就是距离其它所有样本总体距离最小的一个,但是在我们的系统中还考虑了样本的使用频度参数,即在聚类中心附近的几个样本中选取使用频度最大的保留下来,这对保证最终的合成质量起到了一定的积极作用。另一方面,统计出现频率低的样本代表的语境出现频率也低,作为音库容量受限的 TTS 系统,应该优先考虑频繁出现的语境,即为出现频率高的基元分配较多样本。

4.3 音频数据压缩编码算法

我们选择了 GSM 算法对经过裁减的音库进行编码压缩,它的压缩比以及复杂度完全符合嵌入式系统的需求,另一方面可以找到稳定可靠的实现源程序,可以移植到嵌入式系统平台。GSM 算法广泛的应用于手机通讯领域,具有压缩语音质量好,噪声小的特点,压缩比率大概为 1/10。

4.4 音库裁减结果

我们开发了嵌入式汉语 TTS 系统音库管理程序,可以对音库进行切分、裁减、聚类压缩和管理等功能,可以制作指定大小的基于混合基元的音库。例如我们的 TTS 系统音库容量为 300MB,裁减后的音库容量为 70MB,经过 GSM 编码后为 7MB,具体裁减结果如表 3 所示。

表 3 音库裁减结果

基元类型	声母半音节	韵母半音节	音节
原样本数目	29587	29587	23904
原占用空间(MB)	69	112	119
裁减后样本数目	7299	7164	5950
裁减后占用空间(MB)	16	26	28

4.5 听辨实验

经过对 TTS 系统其他模块进行移值后,基于新的混合基元音库的嵌入式 TTS 系统可以在目标平台上流畅的运行。为了检验音库裁减的结果,我们组织了听辨实验针对裁减后的音库和原来的音库进行比较。实验采用 MOS 分法,从新华网上选取了不同题材的 20 句新闻作为听辩材料,分别由裁减前后的两个音库进行合成,将合成的句子打乱顺序由 11 个人进行评分。评分结果为裁减前平均 3.8 分,裁减后平均 3.3 分。由于我们在声韵母的拼接上采用了比较简单的算法,可以看出语音质量有所下降,但是基本上满足了预期的要求。如何有效的改善切分和拼接方法以提高合成语音质量也是我们下一步研究工作的目标。

5 总结与展望

本文采用基于基频包络的音节聚类的方法裁减了音库容量。但是目前的聚类算法复杂度比较高,应该进一步优化以缩短音库的压缩时间。另一方面,将来的工作中还可以考虑基频之外的聚类特征。

为了进一步裁减音库容量,本文首次引入声韵母半音节作为合成基元,实现了基于混合基元的嵌入式语音合成系统。但是目前音节的声韵母切分算法还比较简单,需要提高切分的准确度;声韵拼接平滑算法应当改进以降低或去除产生的噪音;声韵母压缩也可以考虑更多的语境因素和声学参数,这样可以进一步提高合成质量。

嵌入式语音合成系统有着广泛的应用前景,相信随着嵌入式系统应用范围日益广泛,嵌入式语音合成系统将出现在各个领域,为人们提供更加方便的人机交互界面。

参 考 文 献:

[1] 吕士楠. TTS 技术的发展和展望[A]. 第六届全国人机语音通讯学术会议[C],2001.11.

[2] 陈芳等. Natural Sounding Embedded Text-To-Speech Systems[A]. 第五界全国现代语音学学术会议论文集[C],北京,2001.10,302 - 306.

[3] 双志伟,等. 基于量化模版的小型语音合成系统[A]. 第五届现代语音学学术会议文集[C],2001.10,332 - 336.

[4] 孙金城,易立夫. 分层语音合成数据库设计与分析[A]. 全国声学学术会议[C],2002,377 - 378.

[5] Jiawei Han,Micheline Kamber. Data mining : concepts and techniques[M],Higher Education Press,2001.

[6] 吴志勇. 语音基元选取算法及其权重训练[D]. 硕士论文,2001 年,清华大学.

(上接第 8 页)

参 考 文 献:

[1] P. F. Brown, J. Cocke, S. Della Pietra, et al. A Statistical Approach to Machine Translation[J], Computational Linguistics,1990,16(2).

[2] J. Y. Nie, M. Simard, P. Isabelle, et al. Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web[C], 22nd *ACM-SIGIR*, Berkeley, 1999,74 - 81.

[3] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval[A]. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C],2001,105 - 110.

[4] K. L. Kwok, L. Grunfeld, N. Dinstl, et al. TREC - 9 Cross Language, Web and Question-Answering Track Experiments using PIRCS[A]. In: Proceedings of the 9th Text Retrieval Conference (TREC9)[C],2000,419 - 429.

[5] T. Hedlund, H. Keskustalo, E. Airio, et al. UTACLIR-An Extendable Query Translation System. In Workshop on Cross-Language Information Retrieval: A Research Roadmap[C], Organized at 22nd International Conference On Research and Development in Information Retrieval, SIGIR, Tampere, Finland, 2002.

[6] L. Ballesteros, W.B. Croft, Resolving Ambiguity for Cross-Language Retrieval[A], Proceedings of ACM SIGIR[C],1998,64 - 71.

[7] Gao, J. Nie, J Y, Xun, E. Zhang, J., et al. Improving query translation for cross-language information retrieval using statistical models[C], SIGIR 2001,96 - 104.

[8] S. Decker, D. Fensel, F. van Harmelen, et al. Knowledge representation on the web[A]. In: Proceedings of the 2000 International Workshop on Description Logics (DL2000)[C], Aachen, Germany,2000.

[9] I. Horrocks, D. Fensel, C. Goble, et al. The ontology inference layer oil[Z]. Technical report, Free University of Amsterdam, 2000. http://www.ontoknowledge.org/oil/.

[10] 王进,陈恩红,林乐. 一种网络环境中的本体演化和维护模型[J]. 计算机科学,2003,12.

[11] D. Fensel, The role of Ontologies in Information Interchange[A]. In: Proceedings of the 2nd International Scientific and Practical Conference on Programming UkrPROG'2000[C], Ukraine, Kiev, May 2000.

[12] R Neches, R Fikes, T Finin, et al. Enabling Technology for Knowledge Sharing[J]. AI Magazine,1991,12(3):36 - 56.

[13] T. R. Gruber. A translation approach to portable ontologies[J]. Knowledge Acquisition,1993,5(2):199 - 220.