文章编号：1003-0077(2007)02-0094-06

汉语普通话语音合成语料库 TH-CoSS 的建设和分析

蔡莲红，崔丹丹，蔡 锐

（清华大学 计算机科学与技术系，北京 100084）

摘 要：本文介绍了汉语语音合成语料库 TH-CoSS 的建设和分析。本语料库包括男女声朗读语句约 2 万个。语料库分为四个部分：TTS 系统建库用语句、TTS 系统测试用语句、特殊语调语句和特殊音节组。语料设计考虑了语料的平衡和音段、韵律信息的丰富。语料库中除了文本、语音数据外，还带有音段切分标志，标注文件采用 XML 格式。为了方便语音分析与开发，特研制了标注软件。本文还给出了语境特征对语音韵律影响的分析结果。

关键词：计算机应用；中文信息处理；语音合成；汉语；语料库

中图分类号：TP391 **文献标识码**：A

TH-CoSS, a Mandarin Speech Corpus for TTS

CAI Lian-hong, CUI Dan-dan, CAI Rui

(Key Lab. of Pervasive Computing, Ministry of Education, Department of Computer, Tsinghua University, Beijing 100084, China)

Abstract: This paper states our work which focuses on the building and analysis of corpus for Mandarin Text-to-Speech System, named TH-CoSS. The text script consists of four parts: sentences for TTS system building, sentences for TTS system evaluation, special syllable groups, and sentences with special sentence type to convey special intonation. The finished corpus has about 20K sentences read by one female and one male. The annotation files are in XML format, including segmental and prosodic tags. Software tools are developed as well. On the basis of the syllables in TH-CoSS, an analysis of the influences of context features on the prosody of speech is carried out.

Key words: computer application; Chinese information processing; speech synthesis; Chinese; corpus

1 引言

语音合成 TTS(Text to Speech)技术是指将文字转变为语音输出的技术。当前语音合成多采用基于语料库的拼接合成，因此语料库建设在合成系统的设计与实现中扮演重要的角色。建立设计合理、高质量语音语料库具有重要的研究意义和实用价值[1]。

本语音合成语料库（TsingHua-Corpus of Speech Synthesis，简称 TH-CoSS）是由清华大学完成。该语料库面向汉语普通话语音合成的研究、开发和评测。语料本文主要选自新闻，包括汉语普通话男/女声朗读陈述句、感叹句、疑问句，语句长度为5-25个音节。此外语料库还录制了一定数量的轻声、儿化音节组和上声单音节。采样率 16KHz，量化精度 16 位，全部数据约 3GB 字节。语音数据的标注包括汉字、带声调的拼音、音节边界和韵律边界信息，标注文件遵守 XML 扩展标记语言规范。

本文将简要介绍本语料库的设计、标注和分析。其中，第二节和第三节分别介绍语料库的设计和标注，第四节语料库的数据内容，第五节介绍利用决策树分类方法，分析了语境特征对语音韵律影响。

2 TH-CoSS 语料库设计

语料库建设过程包括文本设计、录音、标注等。

收稿日期：2006-05-18 **定稿日期**：2006-07-19

基金项目：国家 863 计划和国家自然科学基金资助项目(60418012,60433030)

作者简介：蔡莲红(1945—)，女，教授，博导，主要研究领域为语音合成和处理，多媒体等。

流程如图 1 所示。在建设过程中，每一步都要进行精心校对，专用的软件工具也是必要的。

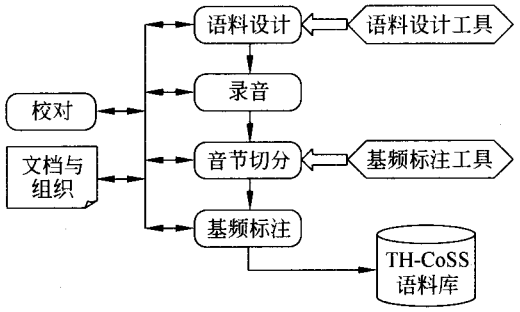

图 1 语料库建设流程

2.1 文本设计

语料库的文本设计是指选取录音文本。这是语料库的关键问题之一[2,3]。涉及的问题包括：明确需求和目的；确定语音基本单位；收集语料素材；研制语料选取算法和软件。目标是最小冗余度、最大覆盖率、科学合理的语料集。在 TH-CoSS 的文本设计中，充分考虑了汉语 TTS 的特殊要求，兼顾语音分析的通用性、标准化，以语音学知识指导的计算机处理。在大量原始语料素材的基础上，针对各子库的不同特点和用途，抽取和构造出各自的录音文本。

2.1.1 TTS 系统建库语料设计

目前大多数汉语 TTS 系统以音节为基本单元，面向新闻播报或信息服务，待合成内容主要是表述事实。因此本语料以陈述语句为主，约5 000 句。此外还包含轻声、儿化词语，以及上声单音节。原始素材取自《现代汉语词典》和大量的新闻报章，并且参考了国家语委公布的必读轻声和儿化词表。

面向汉语 TTS 系统的建库需求，设计原则是兼顾音段和韵律两个层次。先满足音段需求：覆盖汉语的 1 200 多种有调音节。再考虑声调组合、音段音联现象、各种清浊搭配等。在语句挑选时，参照句型分类，考虑句子长度；语句中选录部分轻声和轻读音节；同时以音节为中心，考虑前后音联搭配、音节位置的组合以及音节在自然语流中出现的频率。

首先对原始语料库进行断句和拼音转写，转换成带有拼音的语句、语块（Chunk）或词语。再用 Greedy 算法从大规模语料中选取具有代表性的语句集。这样可得到所需的语料 85％左右。缺少部分由人工设计完成。

文本设计的目标是用尽量少的语料覆盖尽可能多的自然语言现象。这样音库容量就不会太大，又能保证合成语音的质量。初选语料后，再进行校对和优化。校对汉字、拼音、韵律结构；分析统计语音覆盖情况；并修改和增删必要的内容。

2.1.2 TTS 系统测试语料库设计

为了探索合成语音与自然语音的差别，语音质量的主观评价和客观评价方法，系统开发过程中的测试和改进方案，TH-Coss 提供了 1 000 句的测试语料。该部分语句浓缩了不同语境下最常见的语音现象，包括多音字读音、自动分词、常见数字/符号等多方面考察点。可以用于合成语音评价和分析。

普通话 TTS 系统测试语料库的文本原始素材、设计方法与普通话 TTS 系统建库语料库一致，但需注意避免韵律成分的简单重复。不同之处在于，该语料库的设计过程并不是全自动的，部分含有考察点的语句由人工手动添加。显然，这部分语料除用于测试外，也可集成到建库语料中使用。

2.2.3 普通话语调分析用语料库的设计

普通话语调分析用语料库是为普通话语调的分析提供素材，为语调建模、带语气的语音合成提供数据基础。该子库包含疑问句和感叹/祈使句两大部分。与建库语料不同的是，该语料的录制不再限于中性语调，而是要求发音人基于设定的场景进行自然发音。语调分析语料的来源主要是小说、散文等富含情感变化的文字材料。设计时以手工挑选为主，但上下文语境覆盖率仍作为主要的考虑因素。

3 TH-CoSS 的标注

标注是提高语音语料库有效利用的最重要特征。标注系统分为音段标注和韵律标注。而韵律标注和音段标注涉及到标注内容、符号表示、标注文件格式等问题。语音的参考文本也是非常必要的，在本语料库系统中，将其归入文档和标注部分。

3.1 音段标注

音段标注是利用文本标示语音数据的发音。分为“实际发音”和“正则发音”。本语料库选择“实际发音”为标注内容。如：“讲演”标为“jiang2 yan3”。

音段标注设计的基本问题是选择标音符号。目前在基于语料库的汉语拼接合成系统中，多采用音节作为合成基元。因此音段标音可以采用汉语拼音，至于音素单元的读音差别可以从音素的语境信息中获得。如拼音中的“i”分属不同的音节，自然就

可以区分之。考虑到对音段更细致的标注的需要，我们参考MCIPA标音系统[4]，设计了基于26个英文字母和数字集的计算机可读汉语标音符号集作为补充。标音符号采用小写字母表示，音素均用双字母表示，简明规范，计算机处理也很方便。

在静音段部分，借鉴TIMIT的表示方法，把静音段"作为与音段同等的语音单元(sil)来处理，以便前后音联、边界属性等特征的描述和划分"。

3.2 韵律结构标注

韵律标注包括韵律结构标示和韵律参数计算。本语料库一方面根据语音数据，在文本中标注韵律边界信息，另一方面对语音数据进行分析，切分音节，标示音节边界；计算语音的基音频率。

韵律结构标示：音节在韵律结构中的位置是最常用的韵律信息，如音节到韵律边界的距离，边界等级等。TH-CoSS的韵律结构标示为五个层级，并以韵律单元末采样点序号"end_sample"标示韵律边界。自顶向下依次为：

1) 标注的根结构为"utterance"单元，代表整个语音文件中的语音段。

2) "utterance"单元可能被划分为若干个语句，标记为"sentence"。

3) "sentence"单元可能被划分为若干个韵律短语单元，标记为"prosodic_phrase"。

4) 韵律短语单元还可以进一步细分为韵律词单元，标记为"prosodic_word"。

5) 最基本的单元为音节"syllable"。音节的属性，包括汉字(Char)、拼音(Pinyin)、音位音标(MCIPA)、音节结束位置(End_sample)等。另外，静音部分用与音节同层的静音单元加以说明，用"sil"表示。

音节间音联紧密程度标记：在自然语流中，往往出现相邻音节的发音正常却难以切分的情况，在前音节以鼻辅音结尾而后音节以元音开头的情况下尤为常见，吴宗济先生用"你中有我，我中有你"来形容。目前有的合成系统采用混合基元，以提高合成语音的自然度。考虑到这些系统的需要，TH-Coss对"结合紧密"的音节对进行标记"co"，方便了混合基元的使用。此外，为适应TTS建库的需求，还设计了代表语音发音质量的符号。

3.3 TH-CoSS的标注文件格式规范

扩展标记语言XML的简明易懂、可扩展和可移植性强以及结构清晰良好的优点，为标注的结构性、通用性、友好性和可扩展性都提供了有力的保证。

TH-CoSS采用符合xml1.0规范的标注文件格式，在不同操作系统下进行浏览和扩展都很方便，用户可以根据需要添加必要的标注信息，而不用重新设计整个标注体系，只需遵守如下的基本规范。

• 每个声音文件对应一个标注文件——.lab文件。所有.lab的中的标记在一个.dtd文件中定义，注释也全部写在.dtd文件中..lab中心符号全部采用外部定义，且不包括任何注释。

• 标注文件.lab按照语音时间顺序，为树状结构。韵律结构层次定义为元素(ELEMENT)，韵律单元的属性定义为属性(ATTLIST)，根元素为utterance，按照自顶向下的层次，标注文件为树状结构，体现韵律成分的层级关系。

• 所有.lab和.dtd文件严格遵守xml1.0规范。(.lab文件相当于.xml文件，可以在IE上以树状框架显示。)

• 每个声音文件.wav对应的.lab文件记录该段语音中每个音节的汉字、拼音、发音质量和始末位置等信息，以及韵律层级结构。

对于每一个韵律元素，除了末采样点"end_sample"以外的第一个属性是该元素的"标签"属性。它是该元素的最基本描述，在标注工具中将作为标签显示。

3.4 标注工具Visual Speech

为方便数据库的标注与维护，我们还开发了相应的工具软件Visual Speech，工作界面如图2所示。

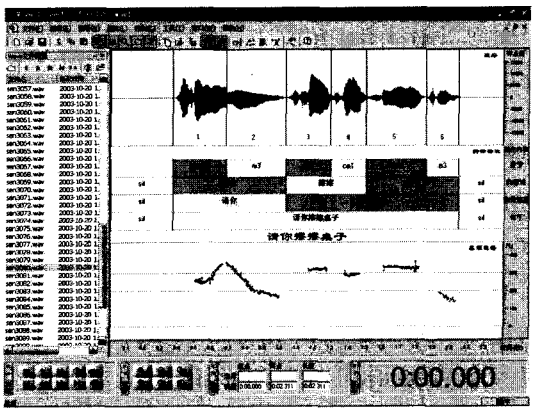

图2 语料库标注工具

其主要功能包括声音文件的显示、修改、更新，音节自动切分，基频、能量、频谱等声学参数高正确

率的自动分析，标注文件的浏览、修改、更新，标注定义的修改与转换，自动文本-音节对齐等。此外，我们还为新的标注内容与自动算法预留了接口。

4 TH-CoSS 的数据内容

设计完成的语料，经过录音、预处理、标注和反复校对，得到如表 1 所示的数据集。

表 2 给出了建库用语料的统计结果。可以看出，其韵律成分的分布较为合理。

录音数据的基本分析：表 3 是语料的语速和基频的分析结果。其中，最大/最小基频是指音节平均基频的最大最小值。可以看出各项数据的分布较为合理，符合中性语调语音的基本特性。系统测试语料比系统建库语速快、起伏变化略大。相比之下，男声比女声起伏变化小。句尾音节的平均基频较低，体现了陈述语句的语调降阶现象。

表 1 TH-CoSS 的数据集

类 别	内 容	规 模
TTS 系统建库语句	陈述句，语句长度 5-25 个音节	5K
测试用语句	陈述句，语句长度 5-25 个音节	1K
特殊音节	汉语轻声音节组	0.8K
	汉语儿化音节组	0.4K
	上声单音节汉语音节表	0.3K
普通话语调分析语料	疑问句，感叹句	1K
语音文件总计		8K
文件	语料库文本(.txt)，语音数据(.wav)，标注文件(.lab)，说明文档：语料库介绍和技术报告(.doc)	

表 2 女声 TTS 系统建库用语料统计

	句子	韵律短语	韵律词	音节数
总数	5 406	16 769	44 658	98 749
平均音节数	18.3	5.9	2.2	—
最大音节数	24	9	4	—

最后还进行了主观听评抽查，随机抽取部分语句进行听评。听音结果表明同一发音人的语音在听感上音质一致、音高稳定、语调自然。而两个发音人的对比听音结果，听音人一致表示男发音人库的语音更为稳定、饱满。在 TTS 系统中应用，合成结果的可懂度与自然度也比较令人满意。

另外，为了方便建立不同规模的语音库，我们还将系统建库语料中的音节按照基频、时长、能量、韵律位置等聚类，提取类中心，在满足音节覆盖的前提下，提取出 1/5 的语音数据，构成建库语料的核心数据集。供 TTS 系统建立小规模数据库，实践表明，合成语音的可懂度满足要求，自然度尚好。

表 3 各部分语料的语速和基频统计

	平均语速(音节/秒)	平均基频(Hz)	最大基频(Hz)	最小基频(Hz)	句首音节平均基频(Hz)	句尾音节平均基频(Hz)
女声建库语料	2.73	233.6	313.0	172.3	279.5	209.9
女声测试语料	3.48	228.0	319.5	171.8	299.1	206.3
男声建库语料	2.78	110.8	161.2	83.2	130.9	101.7
男声测试语料	3.62	108.6	157.2	82.1	128.9	93.1

5 语境特征对韵律表现的影响

语音分析是语音处理中的基础工作，丰富的语料库数据为其提供了物质基础。研究表明，语境特征联系着文本与语音韵律表现。它在语料库文本选取和 TTS 系统基元选取算法中被广泛使用。因此分析语境特征在音节聚类中的作用，可以为基于语

境特征的选音算法提供参考[5]。本文选用决策树工具，通过音节聚类，对语境特征对音节韵律表现的影响进行了分析。给出了不同语境特征对音节韵律表现影响的程度[6]。分析用数据集是TH-CoSS中的主体数据，即TTS系统建库语料。这些数据能比较真实地反映中性风格下，语境信息对自然语音韵律表现的影响。

本文利用CART决策树，用11个语境特征建立问题集，计算韵律参数向量的距离，对TH-CoSS主体数据中的音节进行聚类。分析不同语境特征(问题集)对聚类结果的影响，统计各个特征在决策树中的出现率以及出现的平均层级，并利用一个权重函数，计算每个语境特征对韵律参数影响的权重。语境特征包括：音节在韵律词中的位置(PinW)、在韵律短语中的位置(PinP)、在句子中的位置(PinS)特征。声调音联特征是前音节声调类型(LT)、后音节声调类型(RT)。音段音联特征是：后音节声母(RPh)、前音节韵母(LPh)，后音节声母类别(RType)、前音节韵母类别(LType)、后音节首音类别(RIC)、前音节尾音类别(LFC)。本研究中，韵律参数向量由时长、基频、均方根能量构成。基频向量是该音节长度的0.15、0.5、0.85处的基频平滑值。

决策树生成过程中，问题的出现率及其在决策树中的位置，反映了相应的语境特征对韵律特征的影响，两者分别代表了特征对样本韵律表现区分次数的多少和区分度的大小。统计结果表明：声调音联特征的出现率最高，且后音节声调出现率比前音节的略高。而音节在韵律词或韵律短语中位置的出现率也很高，且出现的平均层级比较低。反映出声调、位置对语音韵律表现的高区分度。

为了综合考察特征在决策分类中的重要性，我们定义了如下权重函数。综合分析特征出现频度及其在决策树中层级。每棵树中特征的重要性权重函数：

$$IF_i(x)=\frac{1}{count(leaf\in i)}\sum_{branch_j\in x}\ \sum_{leaf_k\subset branch_j}\frac{1}{level(k)} \tag{1}$$

其中，$branch_j\in x$ 表示分枝节点 j 属性为 x，$leaf_k\subset branch_j$ 表示叶子节点 k 由分枝节点 j 直接或间接分裂。$count(leaf\in i)$表示树 i 的叶子数。

特征的全局重要性函数：

$$IF(x)=\frac{1}{\sum_i count(leaf\in i)}\times\sum_i(IF_i(x)\times count(leaf\in i)) \tag{2}$$

其中，$count(leaf\in i)$表示树 i 中的叶子节点数，它代表了树的规模。

按照上述函数进行CART聚类，统计特征的出现率和特征出现的平均层级，并计算权重函数 $IF(x)$，结果如表4所示。

表4 建库语料的音节聚类结果

排序 \ 标准	出现率(%)		平均层级		$IF(x)$(%)	
	特征名	统计值	特征名	统计值	特征名	统计值
1	RT	14.85	PinP	2.111 1	PinW	20.52
2	LFC	13.41	PinW	2.255	PinP	16.89
3	LType	13.26	LFC	3.292 4	LFC	15.06
4	LT	12.66	PinS	3.351 1	LType	11.79
5	PinW	11.45	LType	3.765 7	RT	10.81
6	PinP	9.28	LT	4.376 7	LT	9.21
7	RIC	6.70	RT	4.437 8	RIC	4.66
8	LPh	6.53	RIC	4.512 8	RType	4.11
9	RType	6.16	RType	4.667 4	LPh	3.81
10	RPh	2.41	RPh	4.690 5	PinS	1.64
11	PinS	1.88	LPh	5.140 4	RPh	1.47

从该结果可以看出，重要性权重函数综合了特征在聚类决策过程中的出现率和位置信息，综合反映了语境特征对语音韵律表现的影响方式和相对程度。其中，音节在韵律词内的位置被统计出是对韵律表现影响最大的语境特征，韵律短语内位置则位于第二——事实上，某特定音节在韵律结构内的位置很大程度上也影响着前后音联对该音节韵律表现的影响程度和方式。其次是前音节的音段特征，且基于发音部位的分类的权重要高于基于发音方式的分类。然后是声调音联，与音段音联相反，后音节声调的影响略大。统计结果还表明：男女声两部分语料的样本数和类规模相近，两部分语料聚类结果中的特征分布也大致相同，也说明了分析结果具有一定的普遍意义。另外，轻声的分布较为特殊，而对感叹句和疑问句的聚类结果也与陈述句存在差异。

6 结束语

本文介绍了清华大学设计制作的面向汉语普通话语音合成的语料库 TH-CoSS。设计和实现了符合国际通用的 XML 规范的标注系统。本语料库可用于语音合成的开发和评测，也可用于语音分析。目前已提供给国内外多家企业和研究机构。基于决策树的语音分析，研究了语境特征对韵律特征的影响，分析了不同语境特征的相对权重，为新的语料库建设、基元选取提供了有益的参考。

TH-CoSS问世以来，经过反复校对，减少了标音错误、切分误差，满足了用户的需求。随着语音技术的进步，对语料建设提出了新的要求，如自然口语语料、情感语料等。本文提供的思路仍具有较好的参考价值。实际上，语料库建设是一个长期的基础工程。我们也在不断优化设计理念、完善标注系统和软件工具。

参考文献：

[1] 蔡莲红，赵世霞．汉语语音合成语料库的研究与建立[J]．语言文字应用，1999，31(2).

[2] Weibin Zhu, Wei Zhang, Corpus Building for Data-driven TTS Systems [A]. In: Proceedings of 2002 IEEE Workshop on Speech Synthesis [C]. 11-13 Sept. 2002. 199-202.

[3] 孙岭，胡郁，王仁华．中文语音合成系统中的语料库设计[A]．第六届全国人机语音通讯学术会议[C]．深圳：2001.11.

[4] Yiqing ZU, Yingzhi CHEN. A Super Phonetic System and Multi-dialect Chinese Speech Corpus for Speech Recognition [A]. In: ISCSLP [C]. 2002.

[5] Blouin, C., Bagshaw, P. C., Rosec, O.. A Method of Unit Pre_selection of Speech Synthesis Based on Acoustic Clustering and Decision trees [A]. In: ICASSP [C]. 2003.

[6] 崔丹丹，蔡莲红．基于决策树的语料库分析[J]．计算机工程，2006.12.

[7] 蔡莲红，蔡锐．现代语音技术基础与应用[M]．北京：清华大学出版社，2003.