大规模语音语料库及其在 TTS 中应用的几个问题

章 森 刘 磊 刁麓弘

(北京工业大学信息与计算科学实验室 北京 100022)

摘 要 首先介绍了大规模语音语料库以及基于大规模语音语料库的文语转换技术的研究现状,接着介绍了一个大规模连续汉语语音语料库的实例 Slib 的结构和内容;在此基础上,讨论了面向大规模语音语料库的索引技术,提出了语料库检索中的集合运算和最小包容问题,证明了最小包容问题是 NP 完全的,给出了求解该问题的贪婪算法以及算法的近似比;最后,讨论了基于集合运算的大规模语音语料库的检索技术在文语转换系统中的应用,特别是在基本语言单位实例的选取问题上实现了一种基于最小包容的优化方法,对提高文语转换系统的自然度有实用价值.

关键词 语音语料库;集合运算;文语转换;最小包容;信息检索

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.00687

Problems on Large-Scale Speech Corpus and the Applications in TTS

ZHANG Sen LIU Lei DIAO Lu-Hong

(*Information and Computation Mathematics Lab*, *Beijing University of Technology*, *Beijing* 100022)

Abstract The recent advances of large-scale speech corpus (LSSC) and text-to-speech (TTS) technologies are briefly reviewed, then the architecture and annotation information of a large-scale speech corpus Slib are introduced. Based on Slib, the LSSC-oriented indexing methods is discussed, the set operations and the minimum cover problem related to information retrieval in LSSC are presented. The minimum cover problem is a NP-complete problem, and a greedy algorithm is proposed to obtain an approximation solution. The approximation ratio of the proposed algorithm is analyzed. The application and realization of set operations in TTS are presented, and an approach for choosing proper speech instances of linguistic units based on minimum cover is developed, which can improve the naturalness of the synthesized speech of TTS system.

Keywords speech corpus; set operation; text to speech; minimum cover; information retrieval

1 引 言

语音语料库是指为言语技术的研究与开发而建立的语音数据及其标注的集合. 自 20 世纪 80 年代以来,语音语料库的研究、开发与应用在计算机技术的有力支撑下获得了长足进展,先后建立了多种语言的语音语料库,其中的大部分是英语语音语料库①. 例如,英语语音语料库 TIMIT,它是 20 世纪 90 年代初受 DARPA 资助而建立的,包含 630 个说

收稿日期:2008-12-17;最终修改稿收到日期:2009-03-10. 本课题得到国家自然科学基金(60572125)资助. **章 森**,男,1963 年生,博士,副教授,主要研究方向为语音信号处理和自然语言处理. E-mail: zhangsen@bjut.edu.cn. **刘 磊**,男,1980 年生,博士,副教授,主要研究方向为知识获取和智能系统等. **刁麓弘**,男,1980 年生,博士,讲师,主要研究方向为图像处理、模式识别和智能系统等.

① 国际上语料库收集和发布机构: ELRA, the European Linguistic Resources Association; EU Commission's HLT, Human Language Technologies; LDC, Linguistic Data Consortium, http://www.ldc.upenn.edu/; 中文语言资源联盟(ChineseLDC), http://www.chineseldc.org; 国际中文语言资源联盟(CCC), http://www.d-ear.com/CCC/.

话者、5600 多个语句，收集的为朗读语体的美国英语，该语料库对于英语语音识别的研究、开发与评估起到了极大的推动作用，在 Linguistic Data Consortium 提供的多种语音语料库中它被授权、下载的次数最多，至今仍被广泛使用. 但该语音语料库没有标注信息，使用者和开发者需要对其进行标注.

大规模汉语语音语料库对汉语语音处理技术研究和开发的重要性不言而喻. 自 20 世纪 90 年代以来，我国有几十所大学和科研机构相继开展了汉语语音语料库的建设与研究[1-5]. 中国科学技术大学科大讯飞公司发布的汉语语音语料库的规模超过了 2.7Gb 的语音数据，其中包括了男声和女声，采用了手工标注和自动标注相结合的方式，已经应用到科大讯飞的语音合成与识别系统中. 该语音语料库主要面向语音合成，采用了多级标注. 但它是非共享的资源，一般使用者和开发者很难得到授权. 另外，它的标注格式和数据结构是非公开的. 亚洲微软研究院建立的汉语语音语料库包含了大约 18 万个音节，主要应用在了汉语韵律分析以及语音合成中. 该库的标注信息和数据结构是公开的，但标注的信息还不完善. 中国社会科学院语言研究所与多所大学、科研机构合作在 20 世纪 90 年代建立了“863 语音语料库”，收集的是朗读语音. 后来，中国社会科学院语言研究所又相继建立了 CASS 汉语口语语料库等. 中国科学院自动化研究所和声学所、清华大学和北京大学等先后建立了普通话语音语料库，主要用于语音分析与合成的研究[6-7]. 这些语音语料库存在的一个主要问题是标注信息不够多，而且标注的错误率比较高，其原因主要是采用自动或手工标注后的信息没有得到彻底全面的检查校正.

除了大陆之外，台湾、香港、新加坡等地也相继开展了汉语语音语料库的研究[8-12]. 在台湾，中央研究院语言研究所的陈秋豫博士主持建立了大规模汉语口语韵律语料库暨工具平台 COSPRO[11]，它包含 9 个子语料库（约 10.5GB），COSPRO 01-08 属于麦克风朗读语音，COSPRO 09 则为麦克风自发性语音（76MB），内容包罗了不同长度的语料，短至孤立词组（1～4 字词），长至段落语篇（85～996 音节），由 111 位语者因不同的研究目的录制而成，其中包括 61 位女性、50 位男性. 据介绍，COSPRO 语音语料库中的部分语料进行了标注，其标注方法以软件自动标注为主，手工标注为辅，其应用主要面向汉语语音学研究. 该语音语料库的语音数据主要采集自台湾，与大陆的语音在音调和韵律等方面有一定的差别. 另外，中文语言资源联盟（Chinese LDC）、国际中文语言资源联盟（CCC）等都对汉语语音语料库的研究、建设、发布等起到了积极的推动作用.

大规模汉语语音语料库的一个直接应用是文语转换系统或 TTS（Text to Speech），它是利用计算机等平台将文本信息转变为音频数据，以语音的方式播放出来的技术，其追求的目标是让计算机输出的声音清晰、可懂、自然、具有表现力. 文语转换是言语工程和自然人机接口等研究领域的一项关键技术，具有重要的应用价值. 近年来，计算机输出语音的清晰度、可懂度等问题已经基本解决，但计算机输出的语音在自然度、表现力和个性化等方面还存在许多需要研究和解决的问题. 为此，近年来，基于大规模语音语料库的数据驱动和波形拼接技术已成为文语转换技术领域的研究热点，其主要目标是提高和改善合成语音的自然度.

基于大规模语音语料库的数据驱动和波形拼接的文语转换技术的主要问题是如何从大规模语音语料库中快速准确地检索出最合适的语言单位，如字、词、音节或音素等[12-14]. 为此，语音语料库大都采用索引结构，以字、词、音节或音素作为索引项或关键词，以它们在库中的地址（绝对地址或相对地址）作为索引值. 在数据量太大，一级索引不够的情况下，语音语料库还可采用多级索引结构. 索引结构的优势是读取数据速度快，能够满足从大规模语音语料库中快速准确地检索出需要的语言单位的要求，但检索出的语言单位是否最合适？最合适的标准是什么？如何计算和度量？对于这些重要问题近年来发表了一系列的研究报告和论文，其中的主要观点是根据上下文语境以及声学参数和韵律参数等因素作为评价选取的语言单位是否合适的标准和度量，至于如何定义度量公式，如何确定上下文范围，选取哪些声学参数和韵律参数等问题，实现方法差别很大，各有优劣.

本文主要介绍一个大规模汉语语音语料库的架构和标注信息，该库包含了 20 多万个手工标注的音节，而且经过了程序和人工检查校正. 在此基础上，针对大规模语音语料库在文语转换中的应用，提出了基本语言单位的语音实例的最小包容问题，基于集合运算和最小包容实现了大规模语音语料库的快速检索. 上述方法在自主开发的语音合成演示系统中得到了验证，合成的语音非常接近于广播语音.

2　汉语语音语料库 Slib

我们历时两年多的开发，建立了连续汉语语音语料库 Slib，它的语音数据采集自广播语音的真实语料，不含背景音乐和多人的同时语音，说话人 8 个，总数据量约 16Gb，约 110h 的语音数据，其中 2Gb 的语音数据完全采用手工标注的方式进行了多层次标注，标注的 2Gb 语音数据中包括约 22 万个汉语音节，说话人为一个男声和一个女声，各自 1Gb 的语音数据. 语音数据的基本参数是：采样频率 22kHz，16 位量化，单声道，信噪比约在 18～26db 之间，非压缩的 PCM 存储方式.

连续汉语语音语料库 Slib 的建立过程包括：(1) 语音数据采集. (2) 语音数据预处理. (3) 语音数据与汉字对齐. (4) 语音数据与拼音对齐. (5) 时间对齐. (6) 检查修改. (7) 韵律标注. (8) 标注过的语音数据分析. 上述过程中，没有脚本文本的设计，在专用语音语料库，特别是平衡语音语料库中这是重要的一个环节. 由于现在广播和电视的数字化越来越普及，因此语音数据采集主要是在互联网上定向搜集广播和电视的数字化数据，然后进行预处理，包括音频数据的提取、重采样、对语音数据的切分等. 其中，语音数据的切分主要是人工完成，主要包括非语音段和语音质量差的段的剔除，然后按照说话人的 ID 将相应语音段数据存储成语音文件. 这样，采集到的语音数据就生成了约 9600 个语音文件，这些语音文件的大小从几百 Kb 到近 10Mb 不等，每个文件包含了说话人的 ID、生成时间以及语音数据的基本参数等. 语音数据与汉字对齐包括汉字文本标注、分词、词性标注等内容，形成一个标注信息文件，该文件与语音数据文件对应；在标注信息文件的基础上标注拼音信息，采用汉语拼音描述系统；然后标注时间信息，即每个音节的起始和终止时间，再经过程序检查，对查出的错误进行人工修改，形成基本的标注文件；下一步标注韵律信息，最后对标注过的语音数据进行分析和处理.

对语音数据文件标注了汉字、拼音和时间信息后形成了基本的标注文件，该文件与语音数据文件是一一对应的. 因为是手工标注，所以标注文件中不可避免地会存在一些错误，检查修改是必不可少的一个重要环节，否则，可能因为标注错误太多导致语音语料库质量太差，基于这样有错误的语音语料库的统计分析结果是不可靠的. 因为语音语料库规模很大，如果完全依靠人工去检查和修改这些错误，不仅速度慢，而且可能会漏掉一些错误，甚至产生新的错误. 为此，我们采用了程序自动检查和人工修改错误相结合的方法. 对于每一个标注文件，检查程序根据标注规范、标注逻辑检查是否正确，如果发现错误，定位错误并报告错误信息. 其中，拼音正确与否主要根据汉字与拼音的对照词典进行检查，另外还检查声调是否遗漏和正确等；时间检查主要对每个音节的起始时间与终止时间进行逻辑检查，如果发现一个音节的起始时间大于终止时间，或者当前音节的起始时间或终止时间小于上一个音节的起始时间或终止时间，程序就定位和报告错误信息. 根据这些错误报告信息可以快速进行人工修改.

基本标注文件的主要内容包括头部信息和标注信息，其中，文件头部记录了该语音文件的基本参数、说话人 ID 及语音产生的时间等信息. 下面的标注内容是：第 1 列和第 2 列分别记录了每个音节的起始和终止时间，第 2 列也可以省略；第 3 列标注拼音，第 4 列标注汉字并进行了分词，第 5 列标注该行音节的语音质量，共分为 1～9 级，1 级表示最差，9 级表示最好，一般仅标注最差的，其产生的主要原因是音联现象严重、噪声严重等.

语音语料库 Slib 的组织结构是：按照说话人的 ID 为每个说话人建立一个文件目录，在一个文件目录下存放该说话人的所有语音数据和相应的标注文件，其中语音数据文件为 PCM WAV 格式，标注文件为文本格式.

3　倒排索引和集合运算

语音语料库 Slib 中的标注信息包括词、汉字、拼音、起始时间和附加信息等多个层次，其中语音数据约为 2Gb，汉字或拼音共出现约 22 万频次. 据统计，在语音语料库 Slib 中，共有 9600 个语音文件（语音文件大的有十几 Mb，小的有数百 Kb，每个语音文件又伴随一个相应的标注文件，每个标注文件包含的汉字数或拼音数从几十个到数千个不等），出现了 8600 个不同的汉语词（不包括单字词），2900 个不同的汉字，1100 个不同的汉语拼音. 我们的目标是在语音语料库中快速准确地检索各个层次的标注信息，并根据标注信息读取其中的语音数据.

一种常用的方法是对语音语料库中的基本语言单位建立索引，索引项是基本语言单位，索引值是该语言单位在语音语料库的地址或指针. 这里，基本语

言单位或语言单位是指汉语中的词、汉字、拼音或音素等.我们已经指出,这样建立的索引不能满足TTS系统中的最小最大化原则.为此,我们通过建立倒排索引和集合运算解决上述问题,其主要优点是在处理复杂的多关键字查询时,可先将查询的逻辑条件转换为集合的交、并等运算,然后在倒排索引表中得到运算结果后再对语音库中的相关记录进行存取,从而提高查找速度;另一个优点是采用集合运算可以满足文语转换系统中的最小最大化原则,即从最少的语音文件中检索出最多的最合适的语言单位,从而提高合成语音的自然度.

为了叙述方便,我们假设每一个语音文件都有一个唯一的标识符,记为 ID_SF_xxxx;每一个标注文件也有一个唯一的标识符,记为 ID_AF_xxxx,其中后缀 $xxxx$ 表示文件序号,所有标注文件的标识符组成的集合记为 AF.注意,每个基本语言单位都一定出现在一个或多个标注文件中.另外,语音语料库Slib中所有词组成的集合记为 $Slib\_WORD$,所有汉字组成的集合记为 $Slib\_CHAR$,所有拼音组成的集合记为 $Slib\_PY$.下面我们首先介绍几个基本概念.

定义 1. 基本语言单位的伴随集合.如果基本语言单位 X 出现在若干标注文件中,那么,这些标注文件的标识符组成的集合称为基本语言单位 X 的伴随集合,记为 $X\{\}$.

在基本语言单位的伴随集合的基础上,我们可以建立基本语言单位的倒排索引表,如汉语词的倒排索引表、汉字的倒排索引表、拼音的倒排索引表等.据统计,伴随集合的基数从1到数千不等,分布非常不均匀.例如,汉字"的"伴随集合的元素个数近9200,而基数在10以下的占50%多,基数为1的约占20%.

定义 2. 基本语言单位的倒排索引.对于每一个基本语言单位,以基本语言单位作为索引项或关键词,以其伴随集合作为该索引项的值,这样生成的索引表称为倒排索引表,也称倒排索引,其中每一个记录的内容包含两部分,形式如下:

$$X \rightarrow X\{\},$$

其中,X 为一个基本语言单位,$X\{\}$ 为 X 的伴随集合.

因此,对于 $Slib\_WORD$,我们可以建立汉语词的倒排索引表,记为Slib_WORD_IIList;同样,我们可以建立汉字的倒排索引表Slib_CHAR_IIList,拼音的倒排索引表Slib_PY_IIList等.但是,在这些倒排索引表中没有地址指针,只能根据文件的唯一标识符进行存取,效率不高.在存取速度优先的前提下,一般将库中所有的语音数据文件合成为一个或几个较大的语音数据文件,将库中所有的标注文件合成为一个或几个较大的标注文件.在此情况下,倒排索引表中没有地址指针将无法完成数据的存取.因此,我们还需要建立二级索引,即地址索引.

我们以建立标注文件的地址索引表为例,其地址索引表的每一个记录的内容包含两部分,形式如下:

$$ID\_AF\_xxxx \rightarrow ID\_AF\_xxxx\_Address,$$

其中,ID_AF_xxxx 表示文档标识符,而 $ID\_AF\_xxxx\_Address$ 表示该文档在库中的地址.我们记标注文件的地址索引表为Slib_ AF_List;同样,我们可以建立语音文件的地址索引表Slib_ SF_List.因此,对于一个基本语言单位,我们首先在相应的倒排索引表中查找到包含它的文档ID,然后通过地址索引表存取它的相关数据.

定义 3. 基本语言单位的包容.假如给定了一个基本语言单位的集合 $SS=\{S_1,S_2,\cdots,S_N\}$,如果存在库中的 K 个标注文档包含集合 SS 中的所有元素,那么,称这 K 个标注文档组成的集合为集合 SS 的一个包容;如果库中任意 $K-1$ 个或更少的标注文档都不能包含集合 SS 中的所有元素,那么,称这 K 个文档为集合 SS 的一个最小包容.语音语料库中所有标注文档组成的集合记为 SD.

需要指出的是,基本语言单位的包容一定是存在的,但包容和最小包容都可能不是唯一的.例如,对于任意一个基本语言单位的集合 SS,库中所有文档可以构成集合 SS 的一个包容;再如,假设集合 SS 中只有一个元素且该元素的伴随集合的基数大于1,那么,其包容和最小包容都将有多个.

在建立上述索引表的基础上,我们讨论如何利用集合运算实现最小最大化原则.假如给定了一个基本语言单位的集合 SS(事实上,在TTS中一般给定一个句子的文本,然后通过分词等分析算法将其分解为一个个的基本语言单位),而集合 SS 含有的 N 个元素分别为 $S_1,S_2,\cdots,S_N$,即 $SS=\{S_1,S_2,\cdots,S_N\}$.那么,对于任意的 $S_k\in SS$,从倒排索引表中可以找到 S_k 的伴随集合 $S_k\{\}$.通过集合的"交"运算,我们可以得到如下的集合 SS_NI:

$$SS\_NI=\bigcap_{k=1}^{N} S_k\{\}.$$

如果集合 SS_NI 非空,则表明至少有一个文档包含集合 SS 中的所有语言单位.这种情况下,我们找到

了集合 SS 的多个最小包容，但需要一个准则选取哪一个文档，如选取包含基本语言单位最多的文档或者选取语音质量最好的文档等（选取标准可能很复杂）. 如果集合 SS_NI 为空，则表明没有一个文档包含集合 SS 中的所有语言单位. 这时，我们可以通过集合的“并”等运算构造集合 SS 的一个包容，但问题变得比较复杂和困难.

显然，集合 SS 中的所有基本语言单位的伴随集合的“并”SS_NO 是集合 SS 的一个包容，即

$$SS\_NO=\bigcup_{k=1}^{N}S_k\{\},$$

但集合 SS_NO 一般包含的元素太多，我们希望找到一个最小包容. 显然，我们要寻找的最小包容一定是集合 SS_NO 的一个子集，但其子集数目是 $2^{|SS\_NO|}$，其中，$|SS\_NO|$ 表示集合 SS_NO 的基数. 如果采用逐个子集测试的算法，其时间复杂度将是指数级的（相对于问题规模 $|SS\_NO|$ 的增长）. 当 $|SS\_NO|$ 很大时该算法的实现是不现实的. 因此，我们需要寻找其它的可行算法.

首先，对于一些特殊情况，我们可以比较容易地回答上述最小包容问题. 例如，假设集合 SS 中的每个元素的伴随集合两两互不相交，即任意两个伴随集合的交集为空，那么，集合 SS 的最小包容必是从集合 SS 的每个元素的伴随集合中任意抽取一个元素组成的集合，任意一个最小包容中的元素个数都恰好与 SS 中的元素个数相等. 这时，集合 SS 的最小包容的数目是各个伴随集合的基数的乘积，即

$$|S_1\{\}|\times|S_2\{\}|\times\cdots\times|S_N\{\}|.$$

再如，假设集合 SS 中的每个元素的伴随集合两两相交，即任意两个伴随集合的交集非空，而任意三个伴随集合的交集为空. 那么，将集合 SS 的元素的伴随集合两两分组，从每个组中选取一个公共元素可以组成一个最小包容. 这时，最小包容的数目和最小包容中元素的个数都是 $(|SS|+1)/2$ 取整.

对于一般情况，上述最小包容问题是不容易解决的. 其实，定义 3 中提出的最小包容问题是一个 NP 完全问题. 为了叙述方便，我们再用集合的概念将上述最小包容问题描述如下.

定义 4.　假如给定了一个 N 个元素的集合 $SS=\{S_1,S_2,\cdots,S_N\}$，其中 $S_1,S_2,\cdots,S_N$ 也都是集合，令

$$SA=\bigcup_{k=1}^{N}S_k.$$

如果 SA 的一个子集 SAB 满足如下性质：SS 中的每个元素（也是集合）至少包含子集 SAB 的一个元素，则称子集 SAB 为集合 SS 的一个包容；如果对 SA 的任意一个子集 SAC，当 $|SAC|<|SAB|$ 时，SAC 不是集合 SS 的一个包容，那么，称子集 SAB 为集合 SS 的一个最小包容，而子集 SAB 中的元素称为被选中或代表.

在定义 4 中定义的最小包容问题与布尔逻辑中合取范式的可满足性问题（即 SAT 问题）以及众所周知的集合覆盖问题等 NP 完全问题是可以在多项式时间内相互转化的. 因此，最小包容问题也是一个 NP 完全问题.

定理 1.　定义 4 中给出的最小包容问题是一个 NP 完全问题.

证明.　我们继续沿用定义 4 中的符号，定义一元谓词 $P(X)$ 如下：对于 $X\in SA$，当 X 被选中时，$P(X)=1$ 或真；当 X 不被选中时，$P(X)=0$ 或假. 那么，我们可以构造一个合取范式 PP 如下：

$$PP=\bigcap_{S_i\in SS}\bigcup_{X\in S_i}P(X),$$

其中，$\bigcap$ 表示逻辑合取，$\bigcup$ 代表逻辑析取. 根据 Cook 定理，我们可以得出，上述合取范式 PP 的可满足性问题是一个 NP 完全问题. 另一方面，对于定义 4 中给出的最小包容问题，显然任意一个最小包容不仅要满足合取范式 PP，而且还要找出 PP 的成真赋值中选中的代表数最少的一个，即最小包容. 所以，上述最小包容问题是一个 NP 完全问题.

证毕.

定义 4 中给出的最小包容问题与集合覆盖问题也是可以相互转化的. 事实上，如果我们把 SA 的元素看作集合，把 $S_1,S_2,\cdots,S_N$ 看作一个需要覆盖的集合的元素. 再把定义 4 中的包含关系也转换过来，即如果在定义 4 的最小包容问题中，S_k 包含 SA 中的元素 a_{k1}，$a_{k2},\cdots,a_{km}$，那么，在集合覆盖问题中转化为集合 a_{k1}，$a_{k2},\cdots,a_{km}$ 都包含元素 S_k. 这时，集合覆盖问题是如何选取最少的 SA 中的元素（它们被看作集合）并覆盖集合 $SS=\{S_1,S_2,\cdots,S_N\}$. 因此，我们可以利用集合覆盖问题的方法和结论来分析和求解最小包容问题.

根据定理 1，求解最小包容问题，也就是确定一个选取策略：哪些 $X\in SA$ 被选中，哪些不被选中. 另外，对于最小包容问题，我们还有如下的几个相关问题：

(1) 如何求最小包容？

(2) 共有多少个最小包容？

(3) 最小包容中含有多少元素？

对于上述问题(1),我们给出如下的贪婪算法.该算法是多项式时间算法,能够求出一个包容,但不一定是最小包容.

算法 1.

1. 集合 C=空
2. 集合 $E=SA-C$
3. 集合 $F=SS$
4. while F 非空 {

选取 E_k,满足:

$$E_k=\arg\max_{E_i\in F} No\_F(E_i)$$

令$C=C\cup\{E_k\}$

$E=SA-C$

$F=F-\{$集合 F 中含有元素 E_k 的元素$\}$

}

5. 返回集合 C,结束.

下面的例子可以说明上述算法 1 只能求出一个包容,但不一定是最小包容.

例 1. 假设集合 $SS=\{s1,s2,s3,s4\}$,其中

$s1=\{1,2,3,4\}$, $s2=\{1,2,5,6\}$,

$s3=\{6,7,8\}$, $s4=\{7,8,9,10\}$,

那么,SS 的最小包容有 4 个,分别如下:

$\{1,7\}$, $\{1,8\}$, $\{2,7\}$, $\{2,8\}$.

按照上述算法 1,如果第一次找到的 E_k 为 $s2$ 和 $s3$ 的公共元素 6,那么,就只能找到一个有 3 个元素的包容,找不到一个含有 2 个元素的最小包容.

关于求最小包容问题的求解,我们利用图的连通性可以有更直观的解释.对于上述例 1,我们可以作出一个对应的无向图,如图 1 所示,其中图的顶点为集合 SS 中的元素,两个顶点之间有边连接当且仅当这两个顶点作为集合有公共元素,每一个公共元素对应图中的一条边,且在边上标出公共元素.

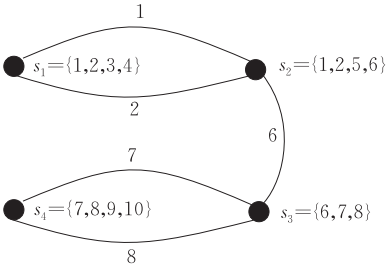

图 1 最小包容(例 1)

根据算法 1,我们首先在最小包容问题的对应图中找出含有相同标号最多的边(不一定是最长路径),然后去掉这些边以及与其关联的顶点,在余下的图中重复上述操作,直到所有的顶点都被去掉,记录下上述操作中每次去掉的边上的标号,这些标号的集合就是一个包容.

下面我们讨论算法 1 的性能问题.我们定义算法 1 的近似比 Ra 为最小包容中含有的元素个数与算法 1 给出的包容含有的元素个数之比.显然,Ra 是大于等于 1 的.根据用贪婪算法求解集合覆盖问题的结论,我们有如下的定理.

定理 2. 求解最小包容问题的算法 1 的近似比为$(\ln N+1)$,其中 N 为集合 SS 的基数.

证明. 因为最小包容问题可以转化为集合覆盖问题,根据贪婪算法求解集合覆盖问题的结论,我们可以得出用贪婪算法求解最小包容问题的近似比为$(\ln N+1)$.详细证明见文献[15].

关于最小包容的另外两个问题,即共有多少个最小包容?最小包容中含有多少元素?本文还没有解决.

4 在 TTS 中的应用

在介绍了大规模语音语料库 Slib 的结构和基于集合运算的最小包容等问题的基础上,我们将讨论它们在 TTS 即文语转换系统中的应用.为了叙述方便,我们称一个基本语言单位(如词语、汉字、拼音等)在某个语音文件中的一次出现为该基本语言单位的一个实例.一个基本语言单位在某个语音文件中可能有多个实例.基于大规模语音语料库的文语转换技术的核心问题是如何从一个基本语言单位的多个实例中选择一个"合适的"实例.这里,"合适的"评价标准不同,选择的策略和结果也不同.

文语转换系统一般包括两个主要模块,即语言处理模块和语音处理模块.语言处理模块的主要任务是分析输入的文本句子,识别出短语、词、专用名称、缩略语、特殊符号等语言单位,将文本句子转换为注音符号表示形式并添加控制符.语言处理模块的输出作为语音处理模块的输入.语音处理模块对于每一个注音符号在语音语料库中查找它的一个合适实例,然后将文本句子对应的所有注音符号的实例拼接起来,在拼接过程中根据控制符的语义对语音实例的音长、韵律特征和停顿时间等进行调整,最后输出一个句子的完整的语音数据流.输出语音的清晰度、自然度等是评价文语转换系统性能的主要技术指标,而自然度是目前最主要的评价指标.

在基于大规模语音语料库的文语转换系统中,其语音处理模块按照处理逻辑可以划分为三层,即数据层、索引层和检索层,见图 2.首先,我们讨论数据层的语音数据的完备性问题.这里,语音数据的

完备是指语音数据能够覆盖所有的基本语言单位的语音实例. 其实，语音语料库的完备性问题涉及语言现象的多个层面，如词语、单字、拼音和声韵母等. 例如，在语音语料库 Slib 中，共出现了 5488 个不同的汉语词(不包括单字词)，约占常用汉语词(约 5 万)的 11%；共出现了 2288 个不同的汉字，约占 GB2312-80 汉字字符集中汉字总数的 33.8%；共出现了 1067 个不同的汉语拼音，覆盖率约 88.55%(汉语拼音约有 1205 个，但各种文献的规定的数目略有不同). 我们看到，在词语、字和拼音 3 个语言层面上，语音语料库都不能 100%覆盖. 因此，在使用语音语料库实现 TTS 之前，必须先进行补充. 语音语料库的补充工作一般仅限于在拼音层面上把所有拼音的语音实例补充完整，而对词语和汉字两个层面不作改变. 根据语音语料库的建立方式，对语音实例的补充方案可以有多种. 一般常用的方案是利用已有的拼音手工拼接出缺失的拼音，然后将补充的所有拼音的语音实例存放在一个语音文件中，再将该语音文件添加到语音语料库中. 在语音语料库 Slib 中，我们也是采用了这种方案.

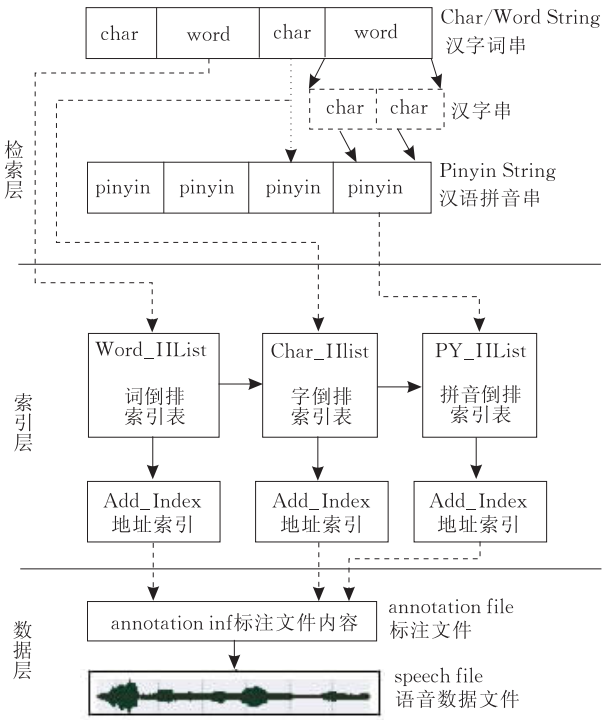

图 2　语音处理模块的三层结构

其次，在索引层建立倒排索引文件和二级地址索引文件. 基于语音语料库 Slib 中的汉语词，建立倒排索引表 Slib_WORD_IIList；同样，建立汉字的倒排索引表 Slib_CHAR_IIList、拼音的倒排索引表 Slib_PY_IIList 等. 对于倒排索引，我们再建立二级索引，主要记录倒排索引文件中的每个文档在语音语料库中的存储地址，以便于存取.

在数据层和索引层的基础上，我们可以在检索层实现对大规模语音语料库的检索和应用. 以文语转换技术为例，在检索层输入的主要是两个字符串：即汉语字词混合的字符串和相应的汉语拼音字符串，这些字符串一般都是以句子为单位的，而期望输出的是该汉语字词混合字符串对应的语音数据流. 在 TTS 中，将输入的汉语字词混合字符串按照顺序依次取出字或词作为基本语言单位，然后在索引层查找是否存在相同的基本语言单位；如果查找成功，则可以在语音语料库中找到该语言单位的语音实例；否则，将词切分为单字，将单字转换为拼音再进行查找. 除非拼音错误，否则总能找到一个拼音的语音实例. 基于语音语料库检索语音实例的逻辑流程如图 3 所示.

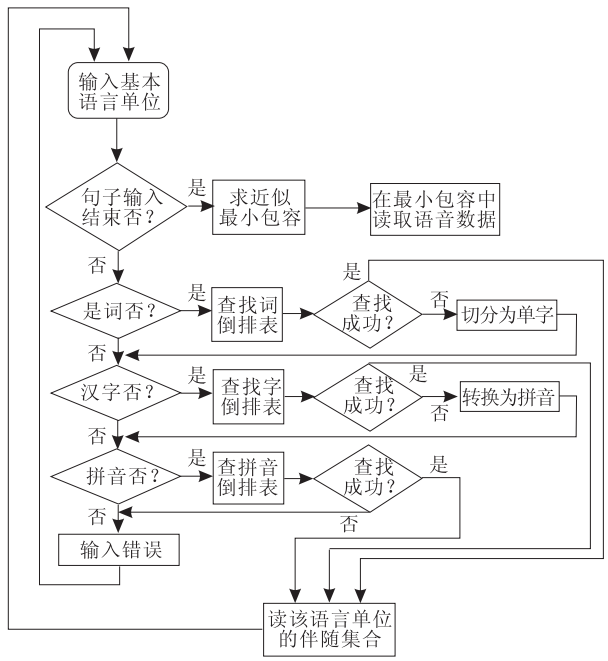

图 3　语音处理模块的检索处理流程

基于语音处理模块的三层结构和检索处理流程，对于每一个输入的基本语言单位，我们可以找到它的一个或多个语音实例，特别是存在多个语音实例的情况下，选择哪一个最合适？这里，“最合适”的评价标准应该是量化的、可实现的、客观的，并且与人的主观评价保持一致性. 在文语转换系统中，一般的选择方法是：针对一个基本语言单位，根据其上下文环境、声学参数、韵律参数等构造一个特征向量 $\boldsymbol{P}$，并对它的每一个语音实例也构造一个类似的特征向量 $\boldsymbol{T}$，计算向量 $\boldsymbol{P}$ 与每一个向量 $\boldsymbol{T}$ 之间的距离，

选择距离最小的特征向量 $\boldsymbol{T}$ 对应的语音实例作为该基本语言单位的语音实例.需要指出的是,在构造特征向量 $\boldsymbol{P}$ 时,可以通过分析输入字符串获取一个基本语言单位的上下文环境,但其声学参数和韵律参数必须通过其它方式获取,如事先通过统计方法获得每个汉语拼音(总共 1200 多个)的音长、基频、能量的均值和极值等信息,并存放在文件中以备查询使用.另外,为简化处理,也可以仅仅使用基本语言单位的上下文作为选择语音实例的匹配条件.

基于上下文选择语音实例是根据当前基本语言单位的左右文本信息作为匹配条件来选择合适的语音实例.虽然上下文信息丰富,但一般仅选择左、右各一个文字或拼音形成匹配规则.这些匹配规则一般都有如下的形式,即左文字+当前基本语言单位+右文字.我们使用了如下的 12 条规则(a)~(l)在大规模语音语料库中选择合适的语音实例.

(a) L_{ch}+WD+R_{ch}

(b) L_{ch}+WD

(c) WD+R_{ch}

(d) WD

(e) L_{ch}+CH+R_{ch}

(f) L_{ch}+CH

(g) CH+R_{ch}

(h) L_{py}+PY+R_{py}

(i) L_{py}+PY

(j) PY+R_{py}

(k) CH

(l) PY

其中,WD 为汉字词,CH 为汉字,PY 为汉语拼音,L_{ch} 为左文字,L_{py} 为左拼音,R_{ch} 为右文字,R_{py} 为右拼音.上述规则的匹配优先级顺序是:从(a)到(l)依次降低.当匹配成功时,表明选择了一个合适的语音实例.

例如,假设文语转换系统的输入是如下的一个汉语句子:

语音处理会议在沪召开

经过语言模块的分析处理,可以得到如下两个字符串:一个是汉语字词混合字符串 CW,另一个是其对应的拼音串 PY,如下所示:

CW:语音/处理/会议/在/沪/召开

PY:yu3/yin1/chu3/li3/hui4/yi4/zai4/hu4/zhao4/kai1

其中,符号"/"为基本语言单位分隔符;对于上述汉语句子中的基本语言单位选择其语音实例的过程如下:

1. 首先取第一个汉字词 WD="语音",其上下文信息为 L_{ch}="sil",R_{ch}="处";在语音语料库中应用规则(a)进行匹配,如果匹配成功,则选择了汉字词 WD="语音"的一个语音实例,否则,再利用规则(b)~(d)依次进行匹配;

2. 如果规则(a)~(d)都匹配失败,那么,从原汉字词的尾部去掉一个汉字形成一个新的基本语言单位 WD="语",这时,其上下文信息为 L_{ch}="sil",R_{ch}="音";再用规则(e)~(g)依次进行匹配;

3. 如果仍然匹配失败,那么,取该基本语言单位对应的汉语拼音和上下文信息,再用规则(h)~(j)依次进行匹配;

4. 如果仍然匹配失败,那么,用规则(k)和(l)依次进行匹配.

因为任一个汉语拼音在语音语料库中至少有一个语音实例,所以,利用规则(a)~(l)和匹配过程(1)~(4)可以在语音语料库中选出一个合适的语音实例.这里,"合适"的标准是根据上述规则(a)~(l)确定的.当语音语料库中的规模较小时,采用上述规则(a)~(l)选取语音单元的前面的规则被命中的概率较小,但占用了较多的查询时间.因此,基于小规模语音语料库的合成一般采用简单的单元选取规则.例如,有些系统仅仅考虑规则(h)~(l)或者(i)~(l).

现在许多合成系统采用仅以拼音为基础的单元选取规则,即从要合成的句子的拼音串出发,按照特定的规则选取语音单元.这种方法的优点是简化了语音单元选取规则,节省了查找时间,但也有不足之处.例如,当汉字到拼音的转换不准确时,选取的语音单元必定出错.如果采用以汉语字、词和拼音为基础的单元选取规则,可以减少上述错误.

本文的语音单元是以汉语字、词和拼音为基础的,并且按照上述规则(a)~(l)顺序执行.例如,汉语句子:"语音处理会议在沪召开",其中语音单元的选取正式按照上述方法完成的.图 4 是本文开发的基于大规模汉语语音语料库的合成演示系统界面,其中右半部分是输出,包括用户输入的句子的处理结果、合成的语音波形显示与播放等.

图 4 语音合成演示系统界面

另外，还有一个系统性能评价问题. 对于本文来说，主要包括两个方面：语音语料库的评价和基于它的合成系统的评价. 前者主要考虑的因素包括语音语料库的规模、字词和拼音的覆盖率等. 本文的语音语料库的规模已经达到 2Gb 语音数据，包括约 22 万个已经标注的汉语音节(还在不断增加)；但由于该库是从真实广播语音中采集的数据，所以存在语料不平衡现象，有些音节存在数百个实例，有些只有一个甚至没有. 这是本库的一个主要缺陷. 对于合成系统的评价主要是评估合成语音的自然度等，大多采用人工试听打分的方法，主观性比较大. 本文中的系统采用了大规模语音语料库，对于库中有相近实例的句子合成出的语音自然度非常高，否则，自然度较差.

5　结束语

本文介绍了大规模语音语料库以及基于大规模语音语料库的文语转换技术的研究现状，并且重点介绍了一个大规模连续汉语语音语料库的实例 Slib 的结构和内容. Slib 的语音数据采集自中央电视台的新闻联播真实语音，手工标注了约 22 万个汉语音节，适合于汉语语音合成与识别的研究与开发. 基于大规模语音语料库查找基本语言单位的语音实例问题是言语工程中的基本问题. 对于该问题，我们讨论了面向大规模语音语料库的索引技术，提出了语料库检索中的集合运算和最小包容问题，证明了最小包容问题是 NP 完全的，给出了求解该问题的贪婪算法以及算法的近似比；最后，讨论了基于集合运算的大规模语音语料库的检索技术在文语转换系统中的应用，特别是在基本语言单位实例的选取问题上实现了一种基于最小包容的优化方法，对于提高文语转换系统的自然度有实用价值. 现在，我们标注的语音语料库的规模还在不断增加，预计 2009 年底将达到 50 万个汉语音节的规模.

参　考　文　献

[1] Sun Ling, Hu Yu, Wang Ren-Hua. Study on the corpus design of a corpus-based mandarin text-to-speech system//Proceedings of the 6th National Conference on Man-Machine Speech Communication (NCMMSC6). Shenzhen, 2001(in Chinese)
(孙岭，胡郁，王仁华. 中文语音合成系统中的语料库设计//第 6 届全国人机语音通讯学术会议论文集. 深圳，2001)

[2] Tang Sheng-Liang, Zhang Shi-Li, Zhang Zhi-Ping, Wu Xi-Hong, Chi Hui-Sheng. Speech-synthesis system based on news broadcasting corpus//Proceedings of the 8th National Conference on Man-Machine Speech Communication (NCMMSC8). Beijing, 2005(in Chinese)
(汤胜良，张士礼，张志平，吴玺宏，迟惠生. 基于新闻联播语料库的语音合成系统//第八届全国人机语音通讯学术会议. 北京，2005)

[3] Wang Tian-Qing, Li Ai-Jun. Design of continuous mandarin speech corpus for automatic speech recognition//Proceedings of the 6th National Conference on Modern Phonetics. Tianjin, 2003(in Chinese)
(王天庆，李爱军. 连续汉语语音识别语料库的设计//第 6 届全国现代语音学学术会议. 天津，2003)

[4] Cai Lian-Hong, Cui Dan-Dan, Cai Rui. TH-CoSS, a mandarin speech corpus for TTS. Journal of Chinese Information Processing, 2007, 21(2): 94-99(in Chinese)
(蔡莲红，崔丹丹，蔡锐. 汉语普通话语音合成语料库 TH-CoSS 的建设和分析. 中文信息学报，2007，21(2)：94-99)

[5] Li Ai-Jun, Yin Zhi-Gang, Wang Mao-Lin, Xu Bo, Zong Cheng-Qing. Research on Chinese annotated dialogue and conversation corpus and phonetics//Proceedings of the 5th National Conference on Modern Phonetics. Beijing, 2001(in Chinese)
(李爱军，殷治纲，王茂林，徐波，宗成庆. 口语对话语音语料库 CADCC 和其语音研究//第 5 届现代语音学学术会议文集. 北京，2001)

[6] Tao Jianhua, Yu Jian, Kang Yongguo. An expressive mandarin speech corpus//Proceedings of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques. Bali Island, Indonesia, 2005

[7] Wu Tian, Yang Yingchun, Wu Zhaohui, Li Dongdong. 2006 MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition//Proceedings of 2006 IEEE Odyssey — The Speaker and Language Recognition Workshop. San Juan, Puerto Rico, 2006

[8] Chou Fu-Chiang, Tseng Chiu-Yu, Lee Lin-Shan. A set of corpus-based text-to-speech synthesis technologies for mandarin Chinese. IEEE Transactions on Speech and Audio Processing, 2002, 10(7): 481-494

[9] Chou F C, Tseng C Y, Lee L S. Selection of waveform units for corpus-based mandarin speech synthesis based on decision trees and prosodic modification costs//Proceedings of the Eurospeech. Budapest, Hungary, 1999

[10] Wang H C, Seide F, Tseng C Y, Lee L S. MAT-2000 — Design, collection, and validation of a mandarin 2000-speaker telephone speech database//Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, 2000

[11] Tseng Chiu-Yu, Cheng Yun-Ching, Chang Chun-Hsiang. Sinica COSPRO and toolkit — Corpora and platform of mandarin Chinese fluent speech//Proceedings of the Oriental COCOSDA 2005. Jakarta, Indonesia, 2005

[12] Tseng Chiu-Yu, Pin Shao-Huang, Lee Yeh-Lin, Wang Hsin-Min, Chen Yong-Cheng. Fluent speech prosody: Framework and modeling. Speech Communication, 2005, 46(3-4): 284-309

[13] Manning Christopher D, Schuetze Hinrich. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 1999

[14] Yi Jon Rong-Wei. Corpus-based unit selection for natural-sounding speech synthesis [Ph. D. dissertation]. Cambridge, MA: MIT, 2003

[15] Cormen Thomas H, Leiserson Charles E, Rivest Ronald L. Introduction to Algorithms. 2nd Edition. Cambridge, MA: MIT Press, 2001

ZHANG Sen, born in 1963, Ph. D., associate professor. His current research interests include speech signal processing and natural language processing.

LIU Lei, born in 1980, Ph. D., associate professor. His current research interests include knowledge acquisition and intelligent system.

DIAO Lu-Hong, born in 1980, Ph. D., lecturer. His current research interests include image processing, pattern recognition and intelligent system.

Background

The work of this paper is supported by the National Natural Science Foundation of China under project "The Study of Non-Linear Features of Speech Based on Re-Producing Kernel" with grant No. 60572125. Up to now, the speech corpus Slib we built contains continuous speech data more than 110 hours, and a small part of it (about 2Gb) was transcribed and annotated by hand which has more than 220000 Chinese syllables. The scale of the speech corpus Slib is large, and its annotation information is precise. Hence Slib can be applied in the research and development of Mandarin TTS and ASR systems. It is a fundamental problem in linguistic engineering to retrieve special speech instances of some basic linguistic units from large-scale speech corpus (LSSC). Regard to this issue, the indexing techniques for large-scale speech corpus is discussed. The set operations and the minimum cover problem related to information retrieval in LSSC are presented. The minimum cover problem is a NP-complete problem. For this problem, a greedy algorithm is proposed to obtain an approximation solution. The approximation ratio of the proposed algorithm is analyzed. The application and realization of set operations in TTS based on LSSC are investigated. Especially, an approach for choosing proper speech instances of basic linguistic units based on minimum cover is developed, which can improve the naturalness of the synthesized speech of TTS system. The scale of the annotated speech corpus Slib is increasing steadily, which is expected to reach 500,000 Chinese syllables by the end of 2009.