多语种语音合成平台的设计与实现*

徐 俊[1+]，蔡莲红[1]，吴志勇[1,2]

[1](清华大学计算机系，北京市 100084)

[2](香港中文大学系统工程与工程管理系， 香港特别行政区)

摘 要: 随着各国交流的不断深入，多语种以及混语种语音合成已经成为信息沟通和人机交互中越来越重要的部分。针对多语种和混语种语音合成的关键问题和现状，本文设计并实现了一个通用并且扩展性很好的多语种语音合成研究平台 THMTTS，并基于其灵活的系统框架提出了一个混语种语音合成的基本流程，对中英日韩四种语言的混语种语音合成进行了语种检测和语音合成的研究。为进一步提高混语种语音合成技术水平提供了可能。

关键词: 语音合成;多语种;混语种;平台;语种检测

1. 引言

语音合成是通过机器将文字转化为声音的技术，以此提供声文并茂的信息表达方式。它通常也被称为文语转换（TTS，Text-to-Speech）。目前，语音合成技术在国际上已经得到了普遍发展，各种语言甚至方言都有其自身的语音合成系统。为了让系统具有更高的重用性、通用性和扩展性，多语种语音合成便成为了国内外研究的热点。

国外对多语种语音合成的研究较早，并且已经设计并实现了不少优秀的多语种语音合成系统。其中比较知名的有贝尔实验室设计并实现的多语种语音合成系统，其流水化的模块结构非常适合做组件测试和评估[1]；还有英国爱丁堡大学的 Festival 系统，它提出一种基于相交关系机制的数据结构来取代传统的多级数据结构[2]，可以方便地将线性、树型等多种数据结构用统一的形式来表示。然而这些多语种语音合成系统都还缺乏一个图形化的实验平台以便研究和观察中间数据；其次，英语或其它西欧语言与中文相比在文本分析方面有很大差别，其框架不能直接应用到汉语语音合成上来；然后，大多数多语种语音合成缺乏考虑混合语种输入的问题。国内的多语种语音合成系统还比较少，大部分仅仅考虑单一语种或者双语种（即中文和英文）的合成。为此，我们设计并实现了 THMTTS 多语种语音合成研究平台，以满足多语种语音合成的研究和应用需求。

* 本工作由国家自然科学基金（60418012），北京市科委项目（H037330010720）资助

* 徐俊, Email: xujun00@mails.tsinghua.edu.cn

2．多语种语音合成平台 THMTTS

2．1 系统结构

设计 THMTTS 的目标在于建立一个可以对多语种语音合成进行研究和分析的平台。通过该平台，一方面可以完成单一语种语音合成，也可以完成多语种或者混语种语音合成，另一方面，可以替换语音合成过程中某一个模块的算法或者算法的具体实现，并能够在平台上观测模块的输出。根据这个设计目标，THMTTS 应当具有下面的功能：

（1）完整的语音合成框架。作为语音合成平台，首先必须要能够完成单语种语音合成的任务，并在此基础上提供多语种和混语种的支持。这个功能一方面需要提供语音合成底层的数据支持，另一方面要提供语音合成过程中与算法无关的工具，比如文件读写，音频播放设备的控制等。

（2）灵活的模块构成。为了达到研究的目的，系统应该能够支持加载和拆卸语音合成过程中的不同算法模块，或者直接使用系统集成的算法模块，以此来进行语音合成中某个过程的算法的研究和分析。

（3）直观的数据显示。为了能够观测和比较算法的效果和模块的输出，一方面需要提供一个图形化的用户界面并支持多种不同数据类型的显示方式，另一方面也必须制定一套较为灵活并且可扩展的数据接口，这样研究人员只要按照接口规范来实现，就可以将其感兴趣的模块输出数据直观的显示在上层的用户界面上。

为了达到通用并且可扩展的目的，多语种语音合成系统首先应当将合成算法和数据分离。此外，数据类型和数据表现也应当分开考虑。因此，THMTTS 的整体结构由 3 个部分组成，分别是：数据类型定义模块、算法流程模块和图形化用户界面模块。模块之间的逻辑关系如图 1 所示：

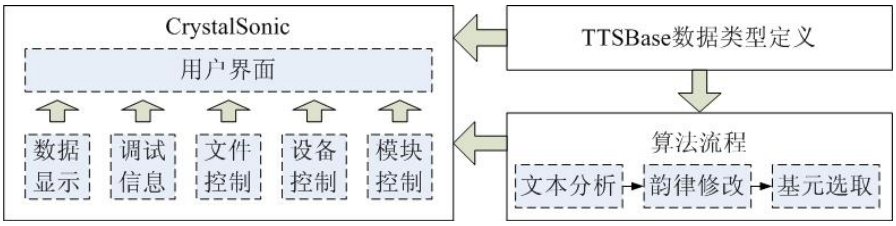

图 1 THMTTS 的系统整体结构

图形化用户界面 CrystalSonic 作为与用户进行交互的入口，它封装了数据显示、调试信息显示、模块加载拆卸的管理、文件读写和字符集编码控制、音频设备的播放管理五个部分。前三个部分是系统设计的要点，其中数据显示的功能提供包括简单数值、字符串、波形数据的直观显示。调试信息则是研究人员在算法过程中输出的信息，按照重要程度分为多个级别。模块控制负责加载所有可用的算法模块并按照配置进行连接。

算法流程模块是由语音合成研究人员自行编写的算法模块的集合，提供了一个框架，然后由研究人员来完成具体的算法。算法的类型和功能都是自由的，只是算法模块之间必须按照预先定义好的接口来实现数据传输，这样才能够使用户界面能够对其进行

管理。

TTSBase 数据类型定义模块提供了一个通用的数据类型定义和数据接口，是整个框架中最为关键的模块。它包含算法内部数据类型的定义、算法模块之间数据流的定义、测试输出信息的定义。这些定义都以通用性和扩展性为首要考虑，不但提供了图形界面从用户自行定义的算法模块中抽取特定数据的方法，也保证了算法模块之间是一个松耦合的状态，方便用户自行设计和实现新的算法。

2．2 系统特性

TTSBase 将语音合成过程中的基本数据结构分为数据（Value）、属性（Property）、单元（Unit）、过程（Process）和语句（Sentence）四类。其中，数据又分成四类：数值、字符串、语音波形和对象指针，前三类数据都可以方便的在图形化界面上显示出来；属性是数据及其名称配对之后的集合；单元是指在具体算法过程中考虑的最小成分，比如词法分析过程的词、语法分析过程中的语法成分；过程就是算法的分析目标，所有的单元都从属于特定的过程；语句即一次语音合成的输入，从语句可以索引到内部所有的过程和属性。假设需要合成“我吃面包”这句话，那么在合成过程中的某个时刻，这些数据结构的逻辑关系可能会如图 2 所示，其中带文字的圈表示单元以及其中的属性。

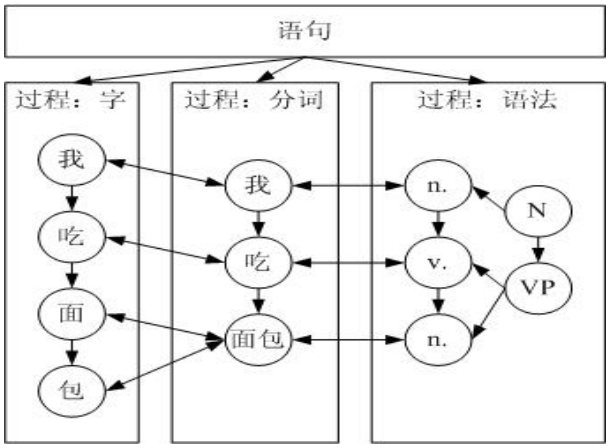

图 2 TTSBase 定义的基本数据结构之间的关系

从图 2 可以看出，单元包含了大部分语音合成过程的数据，其组织形式也体现了算法所使用数据的整体结构。用户界面正是通过调用单元的相关接口来获取数据，供研究和分析。

除了通用的数据结构之外，为了方便研究人员在算法执行过程中实时观察到某些信息， TTSBase 中设计了一个即时信息输出的方法。同时，为了达到信息屏蔽的目的，将调试信息分成几个级别，输出信息的时候只需指定信息的级别，在实际输出的时候，即可根据初始化时的设定来屏蔽掉某些级别的输出。

系统设计的另外一个特性就是算法模块的灵活性。THMTTS 提供了一个通用的模块框架。在此基础上,算法脱离了对数据形式的依赖，研究人员能够集中在算法设计和实现上。另外,每个算法模块都是以动态连接库（dll）的方式实现。平台启动的时候将扫描

指定目录下全部动态连接库，并逐一分析然后加入管理池。之后研究人员可以指定语音合成算法流程需要的算法模块以及其连接顺序。

3．多语种和混语种支持

3．1 基本流程

THMTTS 设计的另外一个目标是实现对多语种和混语种的支持。按照对不同语种的支持方式，语音合成系统的研究分为 4 个类别[3]：单一语种语音合成、简单多语种语音合成、混合语种语音合成、带语种检测的混合语种语音合成。目前，国外已经有不少多语种语音合成的研究，但主要集中在前 2 个类别[4]。THMTTS 的算法模块以灵活和通用为目标，也就是为了方便对不同的语言应用不同的算法流程，达到多语种合成研究的目的。

THMTTS 对混语种语音合成的目标是实现支持中文、英文、韩文、日文的带语种检测的语音合成。混和语种的类型可以分成 3 种[5]：语种 A 的某些词是由语种 B 的词衍生而来，但是遵循语种 A 的构词法；语种 A 的句子中完整的使用了语种 B 的文字和词组，但是依然符合语种 A 的语法；直接在语种 A 中包含语种 B 的文字和词组。韩文和日文含有很多中文字，但是读音和构词与中文已经大不相同，混杂英文的情况更为普遍，因此混合语种的研究是很有必要的。为了达到带语种检测的混语种语音合成的目的，THMTTS 在单语种语音合成的文本分析流程基础[6]上增加了编码控制和语种切分的流程，如图 3 所示。

图 3 混语种语音合成流程概要

3．2 编码转换和语种检测

中日韩三国语言的通用字符编码不尽相同，因此，在进行文本分析的时候需要对输入文本进行文字编码的统一。目前，国际上较为通用的编码为 Unicode，其优点是任何字符占用存储空间大小均为 2 Bytes，这样可以使分析过程中对文本长度的判定非常简单。因此，我们将输入文本统一转换为 Unicode。

语种判定和切分是混语种语音合成中很重要的一部分。由于不同的语种文本分析和韵律模型等方面有很大的差别，因此需要判断输入的语种类型。后续的算法流程可以是同时对多个语种进行处理，也可以分别用不同的算法模块来实现。THMTTS 中的语种判断以语句为单位。判定结果整句都是属于同一个语种，也可以是语句内部混杂了其它语种的词汇。因此，语种判定的过程分为两个步骤：语种检测和语句内部语种片断切分。

语种检测是通过统计概率和的方式根据公式（1）来计算的：

$$p(L_i)=\frac{\sum p(c\in L_i)+p'(L_i)}{\sum[p_0(L_i)+p'(L_i)]} \qquad \text{公式（1）}$$

其中，L_i 表示中日韩英中的某种语言，$p(c \in L_i)$ 为语句中的字符 c 出现在 L_i 的概率，$p(L_i)$ 为判定语种为 L_i 的概率，$p'(L_i)$ 为前一个语句判定语种为 L_i 的概率。用到前一个语句的检测结果的原因是同一个文档中的语句基本语种切换不会太频繁，利用前面的结果进行加权，可以提高当前语句检测结果的准确率。系统对 $p(c \in L_i)$ 预先设定如表 1：

表 1 THMTTS 预定义的文字在语种中出现的概率

$p(c \in L_i)$	字母	汉字	假名	韩语文字
中文	0	1.0	0	0
英文	1.0	0	0	0
韩文	0	0.4	0	1.0
日文	0	0.8	1.0	0

另外，抽取了中文日常会话（可能含有英文词汇）50 句、日语日常会话 30 句（大量使用汉字）、英语日常对话 30 句对算法做了一定的测试。由于没有找到合适的样例韩文的混杂情况没有进行测试。测试结果的正确率分别是 100%、96.7%、100%。其中日语判断错误的一句是因为整句话都是由汉字构成且处于段首。这种情况可以通过在全部判断结束再做一次全局判定来加以修正。

4. 结论

THMTTS 为了多语种和混语种语音合成提供了一个很好的研究平台。它具有良好的人机交互界面，通用的数据结构，灵活的模块构成，有利于语音合成研究人员集中精力对算法进行研究和分析。下一步的工作是对相关接口进一步进行优化，也需要改进平台本身集成的模块算法，并在其上进行情感语音合成和 TTVS（Text-to-Visual Speech）的研究。

参考文献

[1] B. Mobius, et al.“Recent Advances in Multilingual Text-to-Speech Synthesis”. In Fortschritte der Akustik, DAGA’96. DPG, Bad Honnef, 1996.

[2] P. Taylor, et al.“The Architecture of the Festival Speech Synthesis System”. 3rd ESCAWorkshop Speech Synthesis, 1998.

[3] C. Traber, et al. “From multilingual to polyglot speech synthesis”. In Proceedings of the Eurospeech, Budapest, Hungary, September 1999, P835-838.

[4] A. W. Black, et al. “Multilingual Text-To-Speech Synthesis”. In Proceedings of ICASSP, Montreal, Quebec, Canada, May 2004, Volume III, P761-764.

[5] B. Pfister, et al. “Mixed-lingual Text Analysis for Polyglot TTS Synthesis”. In Proceedings of Eurospeech, Geneva, Switzerland, September 2003, P2037-2040.

[6] 蔡莲红等. “现代语音技术基础与应用”. 清华大学出版社，2003年第1版.

Design and Implementation of a Multilingual Speech Synthesis Platform

XU Jun[1+], CAI Lian-Hong[1], WU Zhi-Yong[1,2]

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

[2](Department of SEEM, The Chinese University of Hong Kong, Hong Kong SAR, China)

+ Corresponding author: E-mail: xujun00@mails.tsinghua.edu.cn

Key words: Speech Synthesis; Multilingual; Mixed-lingual; Platform; Language Detect

Abstract: Nowadays, multilingual and mixed-lingual speech synthesis has become more and more important in information communication across different nations. Towards the key problem and current status in such researches, a new multilingual speech synthesis platform - THMTTS - is proposed in this paper. In the first part, the system architecture is presented. THMTTS comprises of 3 parts: basic data structure definition part, which provides a general data structure and information logging mechanism; module definition part, which gives researchers power to design and implement new algorithms for speech synthesis; Crystal Sonic, the graphic user interface (GUI), also the main entry point for speech synthesis, encapsulates the observations for data flow, debug information, module management, as well as handling file I/O and controlling wave-out device. We designed a Multi-level data structure without restricting the contents, and the GUI part is able to call the pre-defined enumeration method to iterate all the data stored and expresses it with different appearances, depending on the data type. Logs are also available to be listed in the GUI, as well as outputting to files or other streams. Another feature of this system is the smart module composition. Modules should implement the same interface and be realized in dynamic linking library (DLL). At the system initialization stage, all the modules stored in the specific place will be loaded, and then, users can manually choose which of them to be used and set the linking order. In the second part, multilingual and mixed-lingual support will be discussed. THMTTS aims to provide speech synthesis with language detection for 4 different languages including Chinese, English, Japanese and Korean. The modular structure itself has advantages for multiple language support. The current system also integrated modules that carry out encoding conversion and language detection. Language detection is based on Unicode, which is a general encoding for international use. The paper also proposed a statistical method based on the sum of probabilities to detect different language, which is proved to be effective by the experiment result. In conclusion, the platform provides general and flexible system architecture for speech analysis and synthesis. Based on this, a basic flowchart for mixed-lingual language detection and speech synthesis is introduced. The proposed architecture makes it possible to improve the quality of mixed-lingual speech synthesis.