计算机语音合成的关键技术及展望

陶建华 蔡莲红
清华大学计算机科学与技术系
智能技术与系统国家重点实验室
tao@tts.cs.tsinghua.edu.cn

计算机语音合成系统又称文语转换系统（TTS 系统），它的主要功能是能将计算机中任意出现的文字，转换成自然流畅的语音输出。它使得计算机不仅能够处理数据，显示图像和文字，还能像人一样的说话，从而使得计算机变得更为亲切、自然。计算机语音合成技术经历了一个飞速发展的过程，目前，已经较为成熟并已大量应用在不同场合，如主页和电子邮件的阅读、文稿校对、人机对话、信息查询等等。本文的目的就是让读者能够了解其一般的工作原理，以及近几年在该领域的研究动态。

1 语音合成系统的组成

一般认为，语音合成系统包括三个主要的组成部份：文本分析模块、韵律生成模块和声学模块。其结构如下图所示。

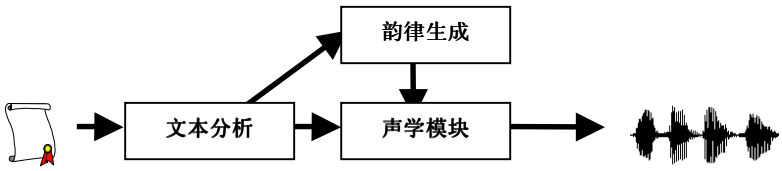

图 1 TTS 系统组成

下面就这三个模块进行较详细地解说。

2 文本分析：

对于语音合成系统来说，首先它要知道要说什么内容。语音合成系统首先处理的是文字。文本分析的主要功能是使计算机从这些文本中能够认识文字，进而知道要发什么音、怎么发音，并将发音的方式告诉计算机，另外还要让计算机知道文本中，哪些是词，哪些是短语、句子，发音时到哪应该停顿，停顿多长等等。其工作过程可以分为三个主要步骤：一、将输入的文本规范化，在这个过程中查找拼写错误，并将文本中出现的一些不规范或无法发音的字符过滤掉；二、分析文本中的词或短语的边界，确定文字的读音，同时在这个过程中分析文本中出现的数字、姓氏、特殊字符、专有词语以及各种多音字的读音方式；三、根据文本的结构、组成和不同位置出现的标点符号，来确定发音时语气的变换以及不同音的轻重方式。最终，文本分析模块将输入的文字转换成计算机能够处理的内部参数，便于后续模块进一步处理并生成相应的信息。其结构如下图所示。

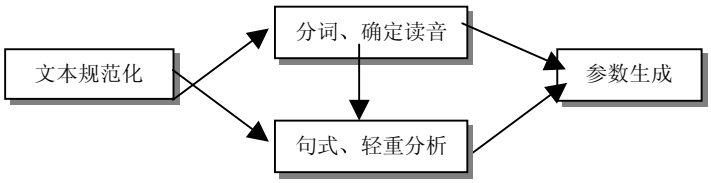

图 2

传统的文本分析，主要是基于规则(Rule-based)的实现方法。这种方法的主要思路是尽可

能将文字中的分词规范、发音方式罗列起来，并总结出规则，依靠这些规则进行文本处理，获得需要的参数。比较具有代表性的有：最大匹配法、反向最大匹配法、逐词遍历法、最佳匹配法、二次扫描法等等。这些方法的优点在于结构较为简单、直观，易于实现，缺点是需要大量的时间去总结规则，且模块性能的好坏严重依赖于设计人员的经验以及他们的相应的背景知识。由于这些方法能够取得较好的分析效果，因此，直到目前，这些方法依然被广泛的使用。

但是近几年来，随着计算机领域中数据挖掘技术的发展，许多统计学的方法以及人工神经网络技术在计算机数据处理领域中获得了成功的应用，让计算机自动从大量数据中提取规律已经完全可能，并已经实现。在此背景下，出现了基于数据驱动(Data-driven)的文本分析方法。具有代表性的有：二元文法法(Di-Grammar Method)、三元文法法(Tri-Grammar Method)、隐马儿可夫模型法(HMM Method)和神经网络法(Neural Network Method)等等。一些比较著名的系统，如 IBM 的语音产品就采用了隐马儿可夫模型法。这类方法的特点是，设计人员根据统计学或人工神经网络方面的知识，设计出一种可训练的模型，并用大量已经存在的数据去训练，将训练得到的模型用于文本分析，而系统设计人员并不需要太强的语言学背景知识。对于工程技术人员，这类方法无疑减轻了他们研究语言学的负担。目前，这类方法在文本分析精度上，已能达到或部分超过了基于规则系统的分析结果，且容易实现多语种的混合，因而被越来越广泛的接受并使用。但是，这类方法也不是没有缺点，它的缺点在于容易使系统获得文本信息的共同特征，而忽略了一些个性。而往往这些个别因素对最终的发音方式影响很大。因此，在也有些系统采取了两者相结合的方法。

3 韵律生成

任何人说话都有韵律特征，比如汉语中，音节有不同的声调，有不同的语气、停顿方式，发音长短也各不相同，这些都属于韵律特征。而韵律参数则包括了能影响这些特征的声学参数，如：基频、音长、音强等。

文本分析的结果只是告诉了计算机发什么音，以及以什么方式发音。这种发音方式还只是抽象的，比如：要发音的声调是二还是三，或是重读还是轻读，到哪应该停顿，而最终系统能够用来进行声信号合成的具体韵律参数，还要靠韵律生成模块。和文本分析的实现方法相类似，韵律生成的方法也分为基于规则的方法和数据驱动的方法。

较早期的韵律生成的方法，均采用规则的方法。这种方法，要求研究人员有大量的音韵学的背景知识，需要对在各种特定的情况下，如声音在句子中不同位置、不同声调、句子的不同语气、甚至是不同的词性下，基频、音长和音强等各个声学参数变化的详细情况，加以总结、归纳。由于各个语种的韵律特征语言相关的，因此针对不同的语种，必须找出与该语种相关联的韵律特征。目前，基于规则的方法，仍然被认作是行之有效的方法。目前大部分汉语的语音合成系统依然采用这种方法。虽然经过研究工作者的努力，这种方法能达到较好的韵律生成效果。但是它也受到很多限制。如前所述，基于规则的方法要求系统设计人员花费大量的时间和精力去研究不同语种的普遍的韵律特征。这是一个非常耗时的工作，且由于规则的复杂性，其生成语音的自然度也受到较多的限制，也就限制了它的一些性能。另外，基于规则的系统方法往往只追求发音的自然，而掩盖了人的个性。如让系统模拟某一个特定人的发音，就显得无力，除非是专门针对不同人而设计一些专用的模型。

目前，通过神经网络或统计驱动的方法进行韵律生成，已获得了成功的应用，目前 Siemens 和 Motorla 公司均采用或试验此一技术。国内，清华大学计算机系在这一方面也进行了大量的研究，其成果也已接近实用阶段。这种方法的实现步骤是：首先设计或收集一个包含大量语音和文本信息的数据，然后建立一个训练模型，用数据库中提取出的韵律参数对模型进行训练，通过训练而得到最终的韵律模型。这种模型的优点，在于保持甚至增强了系统的韵律生成能力的同时，极大的改善了整个语音合成系统的灵活性，便于模拟某一特定人的韵律特征，且为在同一个语音合成系统中整合多语种创造了条件。事实上，有关的研究人员也正在尝试使用这一方法将汉语和其它西方语言整合在一套系统上。

图 3 和图 4，分别反映了基于规则的韵律模型和基于数据驱动的韵律模型，其建立和工作过程。

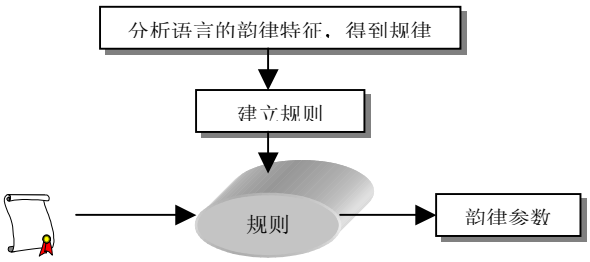

图 3

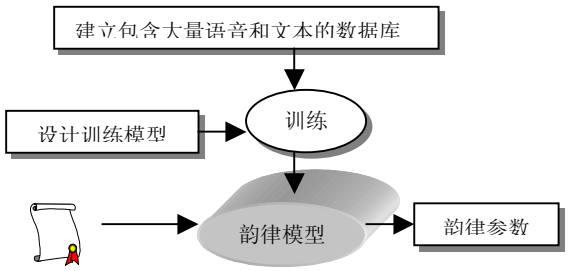

图 4

4 语音生成

知道计算机要说什么以及知道了韵律控制参数后，计算机究竟是怎么来发声的呢，或者它的声音是从哪来的？在系统中，它产生的合成语音是通过一个声学模块来具体实现的。早期语音合成系统的声学模型，多通过模拟人的口腔的声道特性来产生。其中比较著名的有 Klatt 的共振峰(Formant)合成系统，后来又产生了基于 LPC、LSP 和 LMA 等声学参数的合成系统，这些都可以归结为参数合成系统。这些方法用来建立声学模型的过程为：首先录制声音，这些声音涵盖了人发音过程中所有可能出现的读音；提取出这些声音的声学参数，并整合成一个完整的音库。在发音过程中，首先根据需要发的音，从音库中选择合适的声学参数，然后根据韵律模型中得到的韵律参数，通过合成算法产生语音。图 5 反映了它的工作过程。参数合成方法的优点，是其音库一般较小，并且整个系统能适应的韵律特征的范围较宽，但其合成语音的音质却往往受到一定的限制。

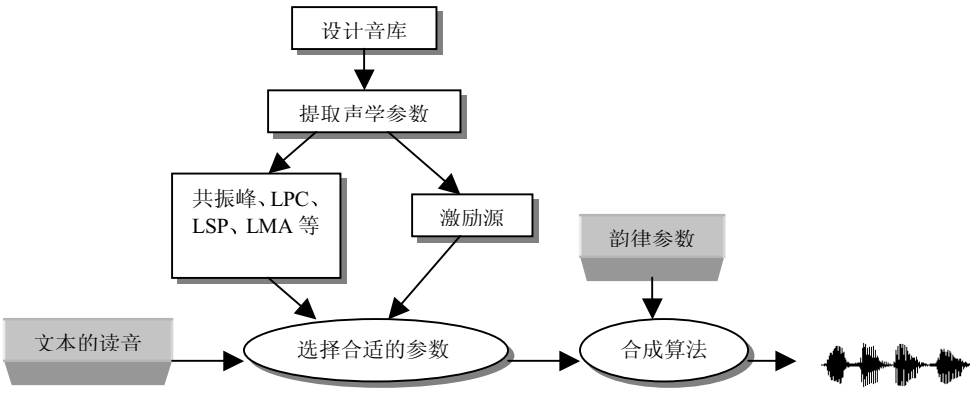

图 5

近十年来采用波形拼接(PSOLA)来合成语音的方法，越来越被广泛的应用。其工作过程如图 6 所示。这种方法的核心思想是，直接对存储于音库的语音运用 PSOLA 算法来进行拼接，从而整合成完整的语音。有别于传统概念上只是将不同的语音单元进行简单拼接，系统首先要在大量语音库中，选择最合适的语音单元来用于拼接，并在选音过程中往往采用多种复杂的技术，包括多项统计学上的技术或神经网络技术，最后在拼接时，使用 PSOLA 算法，对其合成语音的韵律特征进行修改，而使合成的语音达到了很高的音质。如日本 ATR 推出的多语种语

音合成系统，就采用了统计学上的隐马尔克夫模型来进行选音。其它的一些主要语音产品，如 Siemens 的 Papageno 系统，也均采用了类似或相关的技术。

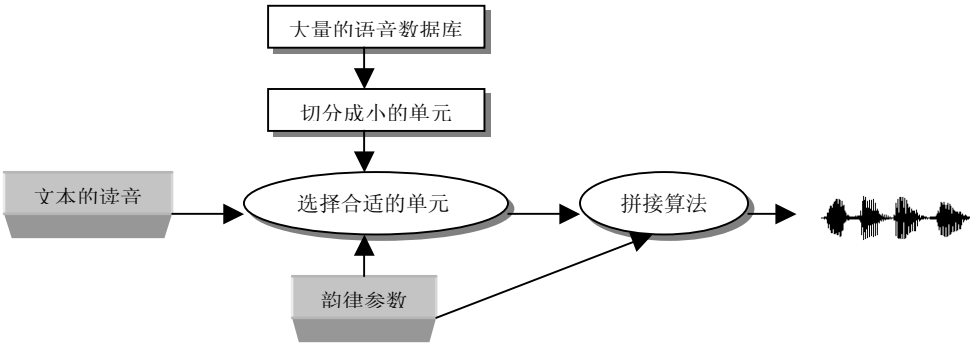

图 6

然而，基于波形拼接方法的系统，也存在一些问题，就是它的音库往往非常庞大，需要占据较大的存储空间。这对系统推广到掌上型电脑或一些小的终端设备上非常不利。另外，在拼接时，两个相邻的声音单元之间谱的不连续，也容易造成合成音质的下降。目前，解决这些问题较好的途径是将两种方法结合起来。在此基础上有诞生了一些新的模型，如基音同步的 Sinusoidal 模型等，这些对进一步改善系统的性能提供了帮助。但目前，这些工作还主要处于研究或实验室阶段。

5 展望

计算机语音合成技术经过近十年的飞速发展，从传统的规则合成技术发展到现在的基于大语料和数据驱动的技术。系统也从单一语种发展到多语种，而且也变得越来越灵活。进一步提高合成语音的自然程度，依然是研究工作者的主要目标之一。目前，其它计算机领域的研究发展，如数据挖掘技术、自然语言理解技术、信号处理技术等，正不断地加速向语音合成领域中渗透，并极大地推动着语音合成系统，朝着能够像人一样，会自然流畅的说话、会学习并能自动模拟的方向发展。