语音合成技术的发展、关键技术及应用

清华大学计算机科学与技术系　　陶建华 博士

目前，“计算机屏幕显示”—这种单调的信息输出方式给用户带来许多不便，特别是在有大量信息输出，且用户又无法直接面对计算机的情况下，如，用户使用的是电话或手机，如何方便有效的获取信息，便显得非常重要。另一方面，长时间地注视显示屏容易使人疲劳，降低人获取信息和理解信息的效率。如果计算机不仅能处理数据，显示图像和文字，还能象人一样的说话，对信息进行讲解，提供声文并茂的信息表示方式，就会让人机之间信息的交流变得更为亲切、自然，以及为超视距获取信息创造更好的条件。因此，语音和语言的研究日益受到重视。语音研究的目的不只是“弥补听官之不足或方便文字之录入”，更重要的是揭示言语交际的机理，获取自然语音中的各种知识和信息，并为人类的信息交流服务。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。它的研究涉及到语音学、人工智能、计算机科学、语言学、心理学等，同时它的研究也推动了相关学科的进步和发展。

一般来讲，实现计算机语音输出有两种方法 一是录音／重放，二是文字—语音转换。若采用第一种方法，首先要把模拟语音信号转换成数字序列，编码后，暂存于存储设备中（录音），需要时，再经解码，重建声音信号（重放）。录音／重放可获得高质量声音，并能保留特定人的音色，但所需的存储容量随发音时间线性增长，而且不能满足实时修改发音内容的需要。

第二种方法是基于声音合成技术的一种声音产生技术。它源于语音生成机理及可计算声学模型。文字—语音转换（TTS）技术是语音合成技术的延伸，它把计算机内的文字转换成连续自然的语声流。若采用这种方法输出语音，应预先建立语音参数数据库、确定语音生成算法等。需要输出语音时，系统按需求先合成语音基元，再按语音学或语言学对自然语言的要求，连接成自然的语流。其特点是，文语转换的参数库不必随发音时间增长而加大。

语音技术已是世界强国竞相研究的热点之一，国内一些科研单位对汉语TTS进行了大量的研究，其中中科院声学所、清华大学、中国科技大学等单位都取得了很好的成绩。目前该项技术已引起了世界上许多著名的计算机厂商或公司的关注，如，L&H、Lucent、ATR、IBM、Microsoft、Dialogic、Siemens和Motorola等，现已研究出多种语言的TTS系统，如汉、英、法、日、德等。其应用领域也在

不断的扩大。

一、语音合成系统的关键技术

1、语音合成系统的构成及工作原理

TTS 系统是用于把文本转换成汉语普通话语音输出的系统，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。其结构如图 1 所示。

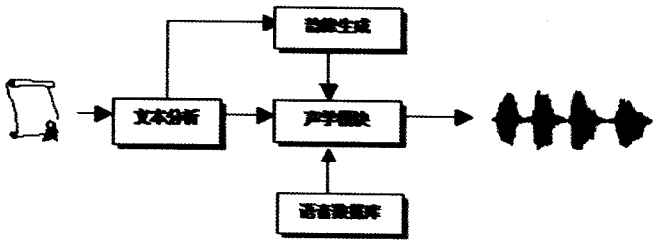

图 1　TTS 的系统框图

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。语音合成系统所处理的最直接对象就是文字，大量的文字堆积在一起就形成文本。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，从而知道要发什么音、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：

①将输入的文本规范化，在这个过程中处理用户可能的拼写错误，并将文本中出现的一些不规范或无法发音的字符过滤掉；

②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、特殊字符以及各种多音字的读音方式；

③ 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。其结构如图 2 所示。

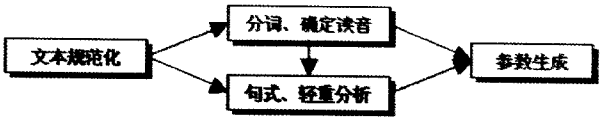

图 2　文本分析流程

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。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，而忽略了特定的词汇或语句的影响，而往往这些因素对最终的结果影响很大。因此，目前比较成功的系统往往采取两者相结合的方法。

韵律生成

任何人说话都有韵律特征，比如汉语中，人说话有四个声调（不包括轻声），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：要发音的声调是二声还是三声，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差别很大，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。虽然有时，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，其生成语音的自然程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，已有一些公司成果采用了或试验了此技术。国内，清华大学计算机系在这一方面也进行了大量的研究，其成果也已进入实用阶段。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。

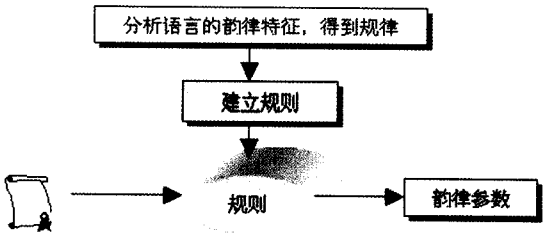

图3　基于规则的韵律模型

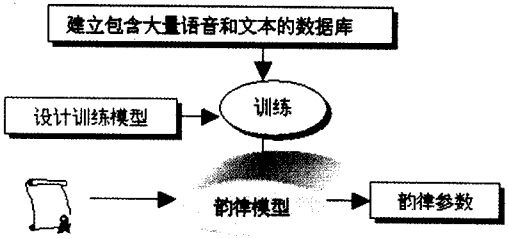

图4　基于数据驱动的韵律模型

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。

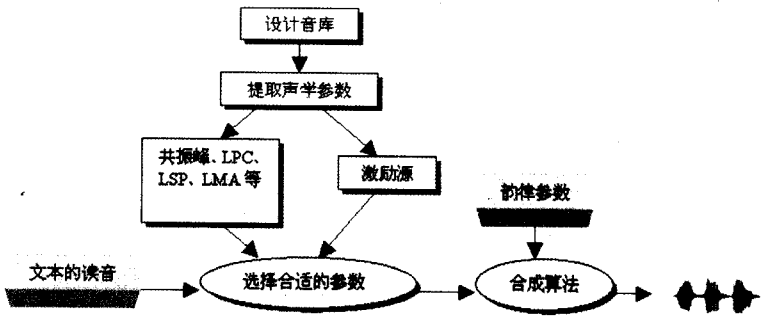

图5　语音生成的矢量运算模型

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。如比较著名的日本ATR推出的多语种语音合成系统，就采用了统计学上的模型来进行选音。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。

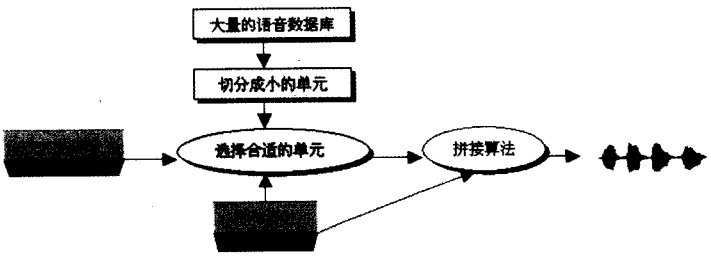

图6 语音生成的波形拼接模型

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。在此基础上也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。

2、说话人模拟和 Talking Head 技术

除了让计算机发出标准的合成语音，说话人模拟，也是语音合成的一个重要研究方向，通过对不同人发音特征、发音音色进行研究，修改语音合成系统中相应的参数，可以达到模拟特定发音人的目的。它可以增加语音合成的个人特色，使语音合成系统具有极大的灵活性和推广性。早期的说话人模拟，通常是直接通过说话人之间的参数映射的方法实现，这种方法，只能实现有限词汇的声音模拟，距离真正的具有个人特色的语音合成系统有着非常大的距离。随着人工智能技术的发展，通过机器学习方法进行语音参数之间的学习和训练，使得语音合成系统按个人特色重新适应变成了可能。

另外，近年来，随着图像处理技术的发展，越来越多的人希望，在语音输出的同时加上人的头像，这样在语音合成系统中加上人的头像，通过头像中的口形与语音的同步动作，将使得计算机语音输出变得更加亲切、友好。Talking Head 技术已经由早期线条勾绘人的基本脸形和口形方法，发展到现在人脸的自然图像进行处理。人脸图像的处理，语音与口形之间同步的研究，也已经变得日趋成熟。这些技术在语音合成里面的应用，极大的推动着语音合成技术本身的技术含量。

3、相关核心技术的研究

语音合成技术不单纯研究语音合成算法本身，实际上是一系列相关技术研究的综合结果，它还包括了语音标注、语音分析、语音分析等相关的基础研究。其中，包括语音的基频分析、共振峰分析、语音的切分和参数提取、韵律标注、声音滤波等一系列的方面的研究。图7展示的是清华大学研制与开发的语音分析和标注工具。它是一个语音研究的综合平台，具有语音切分、基频分析、共振峰分析、能量分析、基频修改、参数提取、语音合成等多种功能。且具有开放的应用接口，目前这个工具已经面向社会，免费发放使用。

二、语音合成系统的应用

当前，信息的“可视”、“有声”、“即时”，可望提高信息的可用性、有效性。减轻工作人员的劳动强度，提高工作效率。文语转换技术应用在电话自动声讯服务中，实现自动应答，专利查询系统、证券代理系统、自动播报系统、Internet信息的有声传送、语音网关（VoiceGateway）等一系列方面，取得了比录音重放更好的适应性和灵活性。本文就语音合成在语音网关及相关领域的应用作一些简要

的描述。

1、VoiceGateway

VoiceGateway语音网关是炎黄新星网络科技公司研发的新一代的语音处理中间系统。其核心技术TTS（Text to Speech）系统采用了清华大学的第二代文语转换技术，支持多语言的语音合成，音速、音质及韵律均可灵活调整，合成质量贴近自然话务播音员。语音网关将TTS包装成一个开放的系统功能模块，用户不需要了解TTS的具体技术细节，通过语音网关提供的开发工具就可以在自己的程序里实现TTS的功能。它支持多种平台，底层通信建立在TCP/IP协议上，可以运行在复杂的网络环境中，同时支持32个并发请求，并通过系统的负载平衡控制，可以在网络中加入任意数量的语音网关系统，把数据处理分配到最佳的语音网关中处理。这种负载平衡控制的方式，理论上可以支持任意数量的系统扩展，甚至可以将语音网关扩展到Internet的任意地方，实现多功能、分布式的语音合成网络。目前语音网关技术已经在互联网、电信领域得到成功的运用。

2、EMAIL TO MOBILE

EMAIL TO MOBILE就是通过用户随身携带的手机，运用语音网关技术实时的“听取”互联网上的电子邮件。每当用户的电子信箱接受到电子邮件时，系统都会自动播打用户注册的手机，通知邮件的到达，用户可根据自己的需要，自己确定是否需要“听取”、恢复或删除邮件。这样的方式使一般人方便了信息获取，减轻了手机屏幕小，需要不断反屏阅读短信息的麻烦。图7是Email到手机的一个工作示意图。

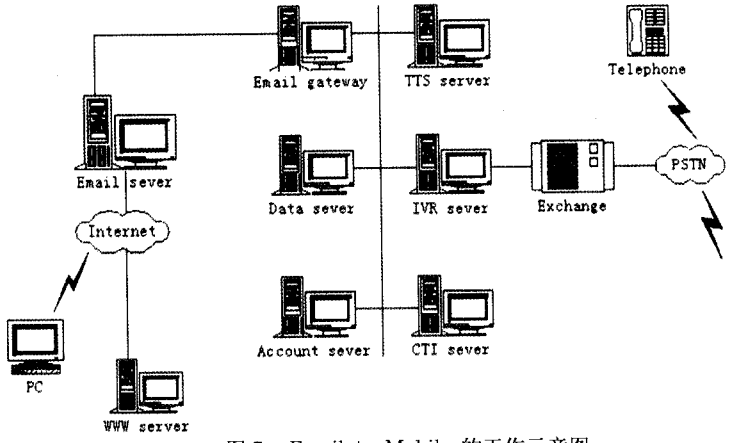

图7　Email to Mobile 的工作示意图

3、统一消息平台

如果说EMAIL TO MOBILE是基于语音合成技术的语音网关的一个具体应用，那么运用语音网关和其他的提供服务的系统共同构建统一消息平台，则极大的扩展了语音技术的应用范围，改变了传统的信息服务模式。在此基础上，可以通过文字、图像、声音的共同方式获取信息，并且用户终端涵盖了互联网浏览器、传真、电话、手机（包括WAP）、寻呼等多种终端设备，而获取的信息也包括了Web、Email以及网络Database等多种信息。在此平台上，可以提供普通用户，声、文、图并茂的综合信息内容，使信息服务变得更友好。图8示例了这种统一消息平台的结构框图。

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

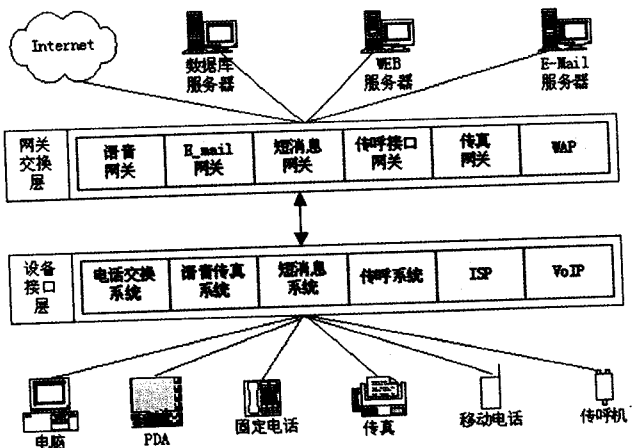

图 8　统一消息平台的结构框图

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。可以预见，在不久的将来，语音合成技术越来越深入有声服务的各个角落，并深入到普通人的生活中。

（上接 48 页）

候。用户可以使用计算机中的麦克风录制问候语，然后把其作为个人问候上传到系统中，也可以使用文本到语音转换器或从任何 wav 源生成问候语。用户还可以收听其当前的问候。

3、独特的平台优势

（1）容量：可以满足数千万名用户的需求。

（2）冗余性：该系统可以实现冗余性，提供了热备份和积极备份功能。

（3）可用性：实践证明，在使用过程中，其硬件可用性高达 99.99%。

（4）灵活的运营商接口：灵活的接口可以配置成与现有的系统一起运行，如用户管 理、计费和告警管理系统。

（5）扩充能力：统一消息容量可以灵活地进行调整，通过提高所需功能的处理能力，满足最大的扩充能力要求。

（6）模块化：通过以模块方式增加功能，可以创建新服务。

（7）支持的标准：特诺漫的统一消息服务支持范围广泛的通信标准，如 Internet 和 WAP，并提供了一条转向 3G 的道路，包括 VoIP（IP 语音）与 WTA（无线电话应用）。

（8）合SMSC：系统会通过短消息通知用户已收到的语音、传真和电子邮件消息。特诺漫提供了一个短消息传输平台，包括通知和其它基于短消息的服务。

特诺漫公司的统一消息有许多优势。第一也是最首要的一点，它可以帮助您在世界上任何地方，以最适合您的格式来获得消息，包括电子邮件、语音信箱、传真邮件或短消息。令使用者拥有一个庞大的信息库，为运营商和供应商提供了一个可以构建一系列其它服务的坚实平台，来创建新型增值服务，确保客户忠诚度、品牌知名度和满意度。