语音合成系统综述及其应用

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

1 概述

随着人机交互技术的进步，计算机将能用自然语言与人类进行交流。为此，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。

人类计算的未来将是无所不在、无所不算。语音计算(Speech Computing) 更是必不可少。它包括对语音的分析、识别、编码、合成、增强等。语音计算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，它的研究将推动相关学科的进步和发展。

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的情况下。长时间地注视显示屏容易使人疲劳，并会降低人获取信息和理解信息的效率。这种枯燥单一的交互方式影响了计算机的应用。如果计算机具备说话的能力, 具有对信息进行讲解的能力，就能提供声文并茂的信息表示方式，可以改变人机交互“默默无闻”的状况，为计算机的普及应用创造更好的条件。

一般来讲，实现计算机语音输出有两种方法：一是录音/重放,二是文一语转换。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/重放可获得高音质声音，并能保留特定人或乐器的音色，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。文字一语音转换（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、发音规则库等。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规则，连接成自然的语流。文一语转换的参数库不随发音时间增长而加大；但规则库却随语音质量的要求而增大。

人在讲话前，首先要有意向(intention)，然后在头脑中形成概念(concept)，最后形成语言。目前，对人类大脑的高级神经活动了解甚少，实现意向到语音的转换有困难，语言合成主要处于文字到语音转换的层次上。

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，其中清华大学、中国科技大学、中科院声学所等单位都取得了很好的成绩。目前世界上已研究出多种语言的 TTS 系统，如汉、英、法、日、德等。Bell 实验室、ATR 和 Siemens 公司研制了多语种 TTS 系统。法国 CNET 实现的多语种 TTS 已在电话网中，用于公共话音服务。

2 语音合成系统的研究现状

人们用语言进行交互时，用声音来表达自己的意向、情感。那么计算机输出的“合成语音”应该是：可懂、清晰、自然、具有表现力。这就是语音合成追求的目标。六十年代首先研制成功英语 TTS 系统。80 年代我国介入汉语合成的研究。九十年代，在国家 863 智能计算机主题的支持下，汉语 TTS 技术有了长足的进步。虽然目前语音合成技术已走向实用，但还

有许多理论和应用问题有待解决。几十年来，专家们构筑了多种多样的 TTS 系统，这里列举一二，以示 TTS 技术的发展和关键。

2.1 从应用的需求出发，设计了特定应用和通用计算机语音输出系统：

- 特定应用的语音输出系统：这种系统适合于特定场合的要求。它可以采用录音/重放技术，或针对有限词汇采用某种拼接技术，不需要语言理解。可用于航班信息发布、语音报时、汽车报站等。
- 文字一语音转换（TTS）系统：这是基于语音合成技术实现的文字到语音的转换是通用计算机语音输出系统。它并不只是文字到语音的简单映射，还包括了对文字的理解，以及对语音的韵律处理。TTS 系统能适应各种应用的需求，应用领域极为广泛。

2.2 从采用的合成技术来分，分为基于规则或拼接合成：

- 基于规则的合成主要是计算参数的轨迹，形成规则，完成语音的参数合成。采用的参数有发音器官参数语音合成：这种方法对人的发音过程进行直接模拟。它定义了唇、舌、声带的相关参数。由这些发音参数估计声道截面积函数，进而计算声波。声道模型参数语音合成：它基于声道截面积函数或声道谐振特性合成语音，如共振峰、LPC、LSP 等参数合成器。这类合成器的比特率低，音质适中，易于实现韵律修改。
- 拼接语音合成技术：它的基本思想存储语音的基元，合成时读取基元、拼接、韵律修饰。拼接语音合成直接把语音基元相互拼接在一起，输出连续语流。这些语音基元取自自然语音的词或句子，它隐含了声调、重音、发音速度变化时的细微特性，合成的语音清晰自然。其质量普遍高于规则合成。但韵律参数修改范围受限。

近年来，规则合成逐渐转向拼接合成。其原因是语音基元的存储不再受限；某些声音如呼吸或爆破音很难由规则合成实现；特别是 80 年代末 E.Moulines 和 F.Charpentier 提出基于波形修改的语音合成算法 PSOLA，拼接合成得到很大的发展与广泛的应用。

PSOLA 就是基音同步叠加。它把基音周期的完整性作为保证波形及频谱平滑连续的基本前提。该算法按以下三步实施：对原始波形进行分析，产生非参数的中间表示；对中间表示进行修改；将修改过的中间表示重新合成为语音信号，由于修改的参数不同，又分为 TD-PSOLA、FD-PSOLA 和 LP-PSOLA。

2.3 基于拼接合成的算法中，需要解决的问题：

- 声学基元的选择：声学基元是指拼接的基本单位。它可能是音素、双音子（Diphone）、三音子（Triphone）、半音节（首音，尾音）、音节、词语、语句等。基元越小，语音数据库越小，拼接越灵活，韵律修饰的规则越复杂。
- 声学基元的样板数：对于同一个基元，由于语境不同，重音表现不同，其声学特征有很大差别。为了减小韵律修饰的负担，可以建立多样板语音数据库。合成时，根据某种规则或模型选择最为理想的基元。
- 韵律修饰：通常 TTS 系统的基元平滑是必须的，其目的是改善合成语音的自然度。而韵律修饰则反映该系统的功能。韵律修饰就是修改语音数据的声学参数，如基频、时长、音量等。通过韵律修饰能力，进行重音、语调的模拟；实现语速、调高的变化。也有的基于拼接合成的系统，通过多样板的选择来体现韵律特征，不含有韵律修饰功能。这就是近年流行的基于数据驱动方法的 TTS 系统。这样的系统中语音数据库非常大，以尽量多的基元样板数来满足韵律的需求。而它的基元选择算法也非常复杂。

3 汉语 TTS 系统的研究

八十年代中科院声学所开始汉语合成的研究。社科院语言所、清华大学、中国科技大学、北方交通大学等单位陆续开展了汉语 TTS 的研究。国外，Bell 实验室、台湾交通大学、台湾大学也研制了汉语 TTS 系统。1999 年，在口语处理国际会议期间，举行了语音合成系统的评比。有十几种语言几十个系统参加。汉语有五个系统。九十年代，国家 863 智能计算机主题资助语音合成的研究，并支持和组织了三次国内汉语语音合成的评测，大大促进了汉语 TTS 技术和系统的进步，并促使语音技术走向实用。

清华大学计算机系人机语音对话研究组成立于 1979 年。1992 年研制成功汉语 TTS 系统 Sonic。经过多年的改进，Sonic 系统可运行在 DOS、Windows、Unix 等操作系统下。该系统的主要功能是：转换国标一、二级汉字为语音。以语音方式输出汉字、词、句子、文章及标点、数字、运算符等。语音输出以句子为单位，按词汇停顿，能自动确定多音字的正确读音。可随时改变声音的幅度（Volume）、基频（Pitch）、速度（Speed）、词间或句间停顿。读出时，可随时“暂停”“恢复”“终止”语音。支持的 DLL，提供文语转换的一系列 API 函数。用户可自行编写发音应用程序。

我们研究了汉语的声调、重音、语调的声学特性，并设计了韵律控制符，进行了重音和语调的模拟。在韵律规则方面，采用了统计和规则相结构的方法研究了汉语韵律规则并进行了一定的韵律模拟工作。在韵律的学习算法方面，我们已经针对神经网络模型中层次结构和训练算法及其输入、输出参数的设计作了一定研究。我们期望通过进一步研究 TTS 系统的韵律学习能力，以优化韵律规则，完善韵律描述，同时改变合成语音千篇一律，缺乏变化的弊病。同时 Sonic 系统获得了多项应用。

4 文语转换的的典型应用

随着计算机技术进入了网络和多媒体时代，语音合成技术也有了飞速的发展，TTS 已应用到信息咨询、电话银行、办公自动化等各个方面。它把声音和文字、图象集成在一起，增强了人们的理解和阅读兴趣，使人与计算机之间的交流变得“亲切”和“友好”。这里列举几个典型的应用。

4.1 电子文档的有声输出

目前，计算机中存有大量的文本，语音合成技术可以提供声音输出，弥补只有屏幕显示的不足。无论以任何方式得到的文字，都可以将其转换成声音。图 1 给出了电子文档有声输出的原理框图。

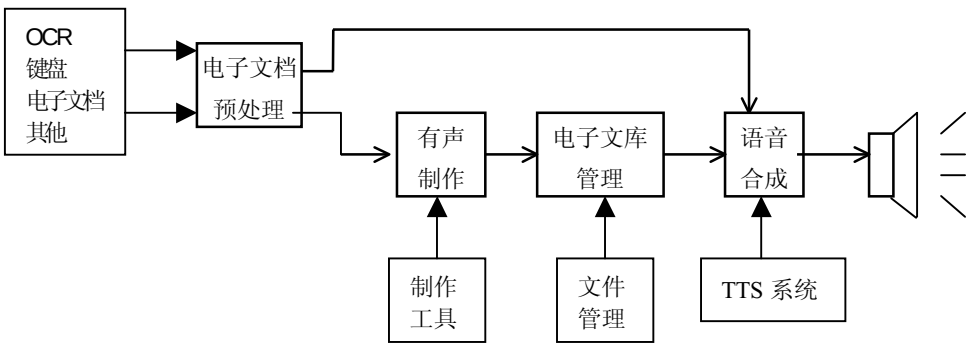

图 1 电子文档有声输出的原理框图

图中的系统包括文本输入、有声制作、电子文库管理、语音合成等几个模块。文档的录入可以多种方式获得，如扫描仪输入，经 OCR 识别后形成文本文件；键盘输入，数据库读出等。一般情况下，电子文档可直接转换成声音。如我们的 TTS 系统和《汉王》公司的汉字识别系统集成，可以读出手写输入或经 OCR 识别后的字符。有声制作模块（不是必须的）的作用是用制作工具给文件加入一系列的标注，以实现语音输出的变化，满足语音输出的特殊要求。如按句、按段改变男/女声、调高和语速等。

4.2 声讯有声服务

网络技术的飞速发展，Internet 服务项目日益增多，时效性提高，电话已成为人与网络交互信息的终端。如通过电话查询股票行情，进行股票交易；或通知或查询即时到达的电子邮件。通过电话进行电子商务活动等等。这里不但用到语音合成技术，还必须与电话技术紧密配合，这就是 CTI（Computer Telephone Integration）或（IT）Internet Telephony。他们促进电话网和数据网的结合，为人们提供了全新的服务。我们曾开发或支持过银行、邮局、劳保、证券、专利、信访声讯服务系统。借助 TTS 技术，把数据库中的文字变成声音，用户利用电话收听即时变化的信息。图 2 给出 IT 解决方案的原理框图。

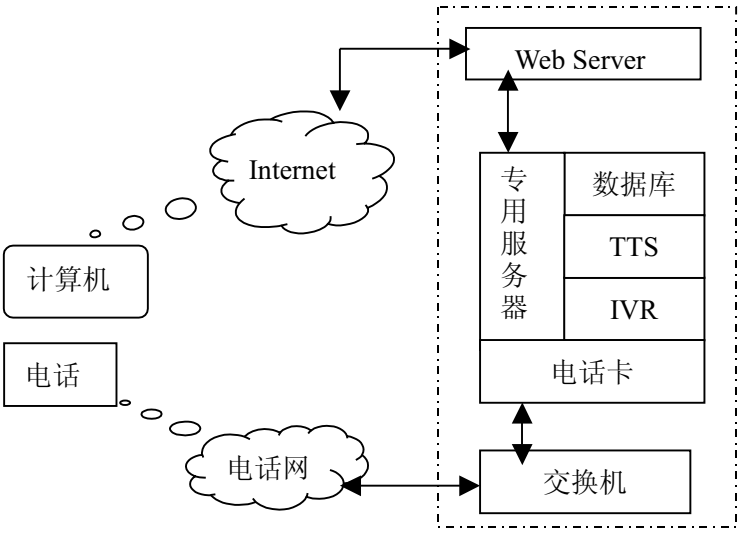

图 2 IT 解决方案的原理框图

用户可以通过计算机或电话得到 Internet 的服务。当以电话方式获得服务时，信息提供商应能自适应地传送用户所需信息。即将文字、图表及有关信息转换成声音，传送到用户的电话上。TTS 技术是实现这些服务的关键。

5 结束语

语音是人机交互的重要手段，具有广阔的应用前景。语音技术已成为智能计算机领域的研究热点，但技术的成熟度、应用的广泛性与需求还有较大的差距。无论从技术的进步、应用的开拓，还需付出巨大的努力。目前，合成语音的可懂度、清晰度基本解决。自然度还不尽人意。表现力差距较大。从应用的角度，把语音输出看作为“锦上添花”是不够的。比如盲人计算机、即时信息服务、语音报警提示、口语机器翻译中语音合成就是非常必要的。语音输出可以为人和电子信息提供声音通道，提高接收信息的速度和效率。，计算机语音技术将伴随我们畅游信息时代。