[文章编号] 1009 - 2145(2002)03 - 0037 - 03

语音合成技术的研究与发展

黄南川[1]　邓振杰[2]　王鬼鬼[3]　张皓健[3]

(1. 河北工业大学廊坊分院，河北　廊坊　065000；2. 华北航天工业学院　计算机科学与工程系，河北　廊坊　065000；
3. 河北工业大学廊坊分院，河北　廊坊　065000)

摘　要：介绍信息技术处理领域的一项前沿技术 ——语音合成技术。简述了语音合成技术的发展历史以及目前国内外在此研究领域的最新成果。讨论了在语音合成技术中用到的一些方法并对这些方法作了简单地分析。阐述了语音合成技术的基本工作原理以及从文字信息到语音输出的工作流程。给出了语音合成技术的系统结构图并对关键的模块作了介绍。简介了语音合成技术的应用领域。对语音合成技术的发展方向，提出了一些看法。

关键词：语音合成；TTS 技术；PSOLA 算法；语言学处理；语音输出

中图分类号：T912. 3　　**文献标识码**：A

Research and Development in Synthetic Technology of the Pronunciation

HUANG Nan - chuan[1]　DENG Zhen - jie[2]　WANG Wei - wei[3]　ZHANG Hao - jian[3]

(1. LangFang College of Hebei Industrial University, Langfang 065000, China; 2. Computer Science & Engineering Department, North China Institute of Astroautic Engineering, Langfang 065000, China;
3. LangFang College of Hebei Industrial University, Langfang 065000, China)

Abstract: It recommends a forward position information disposal technology of the field, the synthetic technology of the pronunciation, sketches out the developing history of the research field and the recent achievements from China and overseas, discusses and analyses briefly the methods used in pronunciation synthetic technology, explain the basic operation principles of the pronunciation synthetic technology and workflow from characters information to pronunciation output, gives out the pronunciation systematic structure charts and recommends the key modules of synthetic technology and the application field. To the developing direction of the synthetic technology of the pronunciation, it also puts forward some views.

Key words: synthetic of the pronunciation; TTS technology; PSOLA algorithm; philology disposal; pronunciation output

1　语音合成技术及发展

在计算机系统中，语音应用技术主要是指基于语音进行处理的技术，主要包括语音识别技术和语音合成技术，是信息技术处理领域的一项前沿技术。

语音识别（SR，Speech Recongnition）技术是指计算机系统能够根据输入的语音识别出其代表的具体意义，进而完成相应的功能。一般的方法是事先让用户朗读有一定数量文字、符号的文档，通过录音装置输入到计算机，于是计算机就准备好了用户的声音样本。以后，当用户通过语音识别系统操作计算机时，用户的声音通过转换装置进入计算机内部，语音识别技术便将用户输入的声音与事先存储好的声音样本进行对比。系统根据对比结果，输入一个它认为最“象”的声音样本序号，就可以知道用户刚才念的声音是什么意义，进而执行此命令。因此通过语音识别技术，计算机可以“听”懂

收稿日期：2002 - 03 - 25

作者简介：黄南川（1959 - ），男，副教授，籍贯河北廊坊，从事计算机应用方面的研究。

人类的语言。

语音合成技术是将计算机自己产生的或外部输入的文字信息，比如文本文件内容、WORD 文件内容等文字信息，按语音处理规则转换成语音信号输出，即使计算机流利地读出文字信息，使人们通过“听”就可以明白信息的内容。也就是说，使计算机具有了“说”的能力，能够将信息“读”给人类听。这种将文字转换成语音的技术称之为文语转换技术，简称 TTS（Text to Speech）技术，也称为语音合成技术。

语音合成技术涉及声学、语言学、数字信号处理技术、多媒体技术等多个领域，是当今世界强国竞相研究的热门技术之一。20 世纪 60 年代英文 TTS 系统首先被研制出来，80 年代我国开始介入汉字 TTS 领域的研究。中科院声学所首先开始汉语合成的研究。之后，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对汉语 TTS 的研究。同时，台湾交通大学、台湾大学和国际上的 Bell 实验室也研制汉语 TTS 系统。近年来，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。清华大学、中国科大、中科院声学所等单位都在这一领域取得了很好的成绩，有些研究成果已经转化为产品得到了实际的应用。如清华大学的 Sonic 系统，中国科技大学的 DK- 863 汉语文语转换系统，杭州三汇公司的中文 TTS 系统，捷通公司的嵌入式 TTS 汉语语音系统，讯飞公司的 KD 2000 汉语文语转换系统等。世界上其它国家也已研究出汉、英、法、日、德等多种语言的 TTS 系统。如 Bell 实验室、ATR 和 Siemens 公司等。法国 CNET 实现的多语种 TTS 已用于电话网中的公共话音服务。1999 年，在口语处理国际会议期间还举行了语音合成系统的评比，有十几种语言的几十个系统参加，其中有 5 个是汉语系统。

2　语音合成技术的方法

语音合成技术可分为参数合成和波形拼接两种方法。早期的研究主要是采用参数合成方法，它是计算发音器官的参数，从而对人的发音进行直接模拟。如著名的 Klatt 的共振峰合成系统。在汉语语音合成方面，研究人员研制出了一些基于共振峰模型的应用系统。如社科院语言所的 SIFS 合成器、中科院声学所的 KX1 系统中基于 Holmes 的并联型共振峰合成器模型，而同样由中科院声学所开发的第二代共振峰合成器 KXFSS 则基于 Klatt 合成器。由于准确提取共振峰参数比较困难，虽然利用共振峰合成器可以得到许多逼真的合成语音，但是整体合成语音的音质难以达到文语转换系统的实用要求。

因此后来又产生了基于 LPC、lSP 等声学参数的合成系统。LPC 合成技术的优点是简单直观，对于单个合成基元来说能够获得很高的自然度。LPC 合成技术是一种时间波形的编码技术，从本质上来说只是一种录音加重放，对于合成整个连续语流，LPC 合成技术的效果是不理想的。自 20 世纪 80 年代末期至今，语言合成技术又有了新的进展，特别是基音同步叠加（PSOLA）方法的提出(1990) 使基于时域波形拼接方法合成的语音的音色和自然度大大提高。PSOLA 技术的主要特点是：在拼接语音波形片断之前，首先根据上下文的要求，用 PSOLA 算法对拼接单元的韵律特征进行调整，使合成波形既保持了原发音的主要音段特征，又能使拼接单元的韵律特征符合上下文的要求，从而获得很高的清晰度和自然度。

20 世纪 90 年代初，基于 PSOLA 技术的法语、德语、英语、日语等语种的文语转换系统都已经研制成功。这些系统的自然度比以前基于 LPC 方法或共振峰合成器的文语合成系统的自然度要高，并且基于 PSOLA 方法的合成器结构简单易于实时实现，有很大的商用前景。

最近几年，一种新的基于数据库的语音合成方法正引起人们的注意。在这个方法中，合成语句的语音单元是从一个预先录下的庞大的语音数据库中挑选出来的，不难想象只要语音数据库足够大，包括了各种可能语境下的语音单元，理论上讲有可能拼接出任何语句。由于合成的语音基元都是来自自然的原始发音，合成语句的清晰度和自然度都将会非常高。

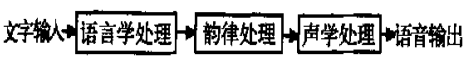
目前，主要的语音合成技术是共振峰合成技术和基于 PSOLA 算法的波形拼接合成技术。这两种技术各有所长，共振峰技术比较成熟，有大量的研究成果可以利用，而 PSOLA 技术则是比较新的技术，具有良好的发展前景。过去这两种技术基本上是互相独立发展的，现在许多学者开始研究它们两者之间的关系，试图将两者有效地结合起来，从而合成出更加自然的语流。例如清华大学的研究人员进行了将共振峰修改技术应用于 PSOLA 算法的研究，并用于 Sonic 系统的改进，研制出了具有更高

自然度的汉语文语转换系统。

随着人们对语音合成的自然度和音质的要求越来越高，PSOLA 算法表现出对韵律参数调整能力较弱和难以处理协同发音的缺陷。因此，人们又提出了一种基于 LMA（对数振幅近似）声道模型的语音合成方法。这种方法同传统方法相比，具有音质好，对时长和声调适应性强，可以灵活调节韵律参数等优点。比较好的解决了 PSOLA 算法难以解决的协同发音问题，因此具有比 PSOLA 算法更高的合成音质。

3　语音合成技术的基本结构

TTS 的基本结构可分为语言学处理、韵律处理和声学处理三大模块。工作流程如下图所示。其基本工作原理是：事先将全部的汉语音节进行录音，形成音频数据，以音库的形式存放在计算机的磁盘上，以供调用。然后用键盘、光电扫描等输入手段，形成 ASCII 文本文件（最新的系统也允许为 WORD 文件、INTERNET 文档，如博欣文公司的《电脑播音员》）存放在磁盘上。系统运行时，先将 ASCII 文本文件进行语言学处理、韵律处理，得到语流控制参数。然后读取音库，从音库中得到对应的音频数据，再经声学处理形成连续的语声流，即完成了从文本到语音的转换过程。



3.1　语言学处理

语言学处理在文语转换系统中起着重要的作用，主要模拟人对自然语言的理解过程，使计算机对输入的文本能完全理解并给出后两部分所需的各种发音提示。其工作过程可以分为三个主要步骤：

(1) 文本规整

将输入的文本规范化。在这个过程中，要查找拼写错误，并将文本中出现的一些不规范或无法发音的字符过滤掉。

(2) 词的切分

分析文本中词或短语的边界，确定文字的读音，同时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。

(3) 语法分析和语义分析

根据文本的结构、组成和不同位置上出现的标点符号，确定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。

3.2　韵律处理

为合成语音规划出音段特征，如音高、音长和音强等，使合成语音能正确表达语意，听起来更加自然。韵律处理有基于规则和数据驱动两种方法。

3.3　声学处理

根据前两部分处理结果的要求输出语音，即合成语音。

4　TTS 技术的应用与发展方向

TTS 技术已广泛用于电子文档的有声输出和声讯有声服务。例如：(1) 金融：帐目查询、交易委托；(2) 邮电：话费查询、话费催缴；(3) 航运：货运查询、客运查询、票务处理；(4) 政府：税务催缴、工商服务；(4) 企业：语音信箱、工业遥控；(5) 教育：高考咨询、辅教服务；(6) 信息：中介服务、商情通告。

TTS 将在下面几个方向发展：(1) 提高语音合成的自然度，达到更加流利和自然的程度。(2) 丰富合成语音的表现力，使得 TTS 技术可以实现各种音色（包括不同性别、不同年龄等）的语音输出。(3) 解决中文与其它语种混读问题。(4) 实现多语种的语音合成，即实现方言、少数民族语言的合成技术。(5) 降低语音合成技术的复杂度，减少音库容量，扩大应用领域。(6) 与网络技术相结合。(7) 可视化的语音合成技术。(8) 为各行业提供 TTS 核心技术和解决方案，特别是 CTI 和嵌入式系统。

可以预料，随着 TTS 技术的进步和 TTS 与其它各种新技术的相结合，语音合成技术必将在更为广泛的范围内得到推广和应用。

参考文献：

[1] 黄南川，罗　恒，蔡莲红，等. 谈谈电话语音系统 [J]. 中国计算机用户，1996，9 (17)：57 - 59.
[2] 王仁华. 让计算机开口说话 [J]. 计算机世界报“产品与技术版”，2000，10.
[3] 唐　浩. 语音合成技术应用实例 [J]. 计算机世界报“产品与技术版”，2000，10.