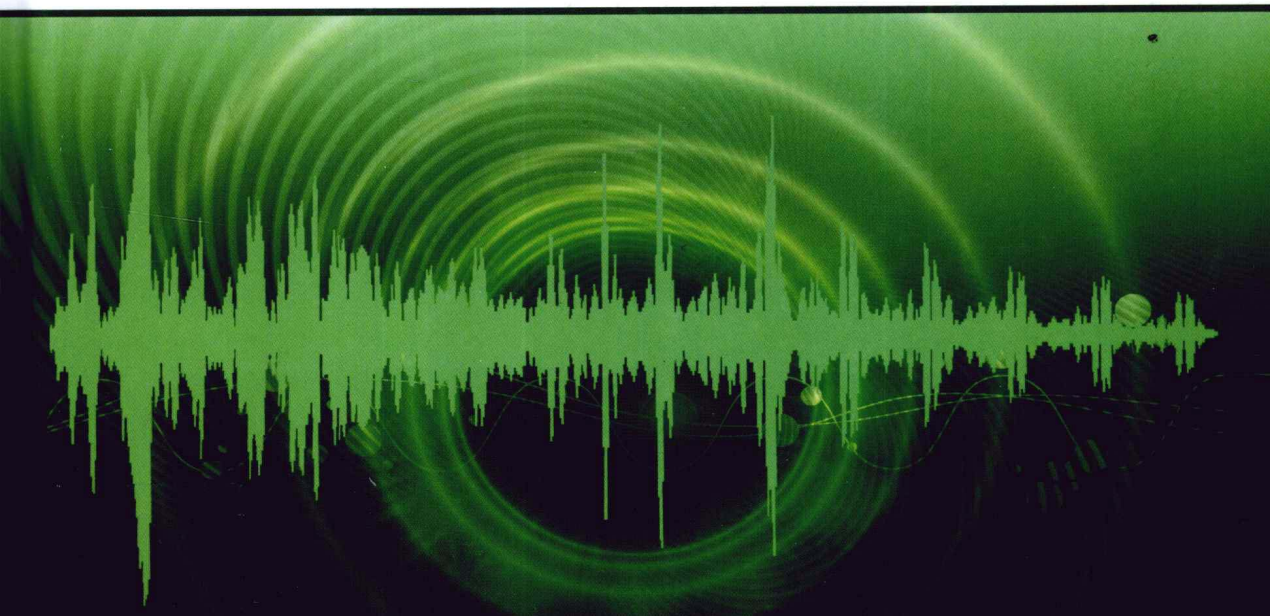
“十二五”国家重点图书出版规划项目

现代声学科学与技术丛书

汉语语音合成

——原理和技术

吕士楠 初 敏
许洁萍 贺 琳 著



“十二五”国家重点图书出版规划项目

现代声学科学与技术丛书

汉语语音合成

——原理和技术

吕士楠　初　敏　许洁萍　贺　琳　著

科学出版社

北京

内 容 简 介

本书介绍语音合成的原理和针对汉语的各项合成技术，以及应用的范例. 全书分基础篇和专题篇两大部分. 基础篇介绍语音合成技术的发展历程和作为语音合成技术基础的声学语音学知识，尤其是作者获得的相关研究成果(填补了汉语语音学知识中的某些空白)，并对各种合成器的工作原理和基本结构进行系统的阐述. 专题篇结合近十年来国内外技术发展的热点和方向，讨论韵律分析与建模、数据驱动的语音合成方法、语音合成数据库的构建技术、文语转换系统的评估方法、语音合成技术的应用等.

本书面向从事语言声学、语音通信技术，特别是语音合成的科学工作者、工程技术人员、大学教师、研究生和高年级的大学生，可作为他们研究、开发、进修的参考书.

图书在版编目(CIP)数据

汉语语音合成：原理和技术/吕士楠等著. --北京：科学出版社，2012

(现代声学科学与技术丛书/田静，程建春主编)

“十二五”国家重点图书出版规划项目

ISBN 978-7-03-032920-2

Ⅰ. ①汉… Ⅱ. ①吕… Ⅲ. ①汉语-语音合成 Ⅳ. ①H11

中国版本图书馆 CIP 数据核字 (2011) 第 247852 号

责任编辑：刘凤娟 / 责任校对：包志虹

责任印制：钱玉芬 / 封面设计：王 浩

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京佳信达欣艺术印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2012 年 1 月第 一 版 开本：B5(720×1000)

2012 年 1 月第一次印刷 印张：24 1/2

字数：470 000

定价：88.00 元

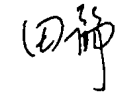
(如有印装质量问题，我社负责调换)

序　一

语言声学是我国声学研究长期坚持的重要领域之一. 早在 20 世纪 40 年代, 马大猷教授就有《国语中语音分配》的论文发表. 新中国成立后, 马大猷先生在筹备组建中国科学院声学研究所的过程中, 就专门设立了从事语言声学研究的部门, 并绵延至今、人才辈出、成果丰硕, 在国内外产生了极大的学术影响, 曾获多项国家级奖励, 为我国语音识别、语音合成和语音通信领域的技术发展和高级人才培养做出了突出的贡献. 《汉语语音合成 —— 原理和技术》一书就从一个侧面反映了相关的研究和应用进展.

该书以独特的视角审视语音合成技术的意义, 把包括语音识别和语音合成的人机对话系统看作人类文明发展史上的重要里程碑. 该书介绍了语音合成的历史和汉语语音合成发展历程; 对作为语音合成基础的声学语音学知识, 如言语声的产生、声音在声道中转播、声道的共鸣以及言语声的感知等做了系统的阐述, 特别强调了合成语音的韵律控制. 作者在综述吴宗济、沈炯等同行的研究成果的基础上, 重点围绕汉语的韵律特征及其对合成语音自然度的影响, 开展了深入的研究, 弥补了汉语韵律知识方面若干空白, 独立构建了汉语合成的韵律模型. 该书的基础篇重点介绍了两款共振峰语音合成器的工作原理, 以及所开发的汉语语音合成器 KX-1、KX-2 和 "联想佳音" 汉语文语转换系统的实用化产品. 在专题篇中, 介绍了基于大规模语音库的波形拼接合成技术. 作者通过英特尔、微软、摩托罗拉和捷通华声公司等企业对该技术的进一步完善及应用, 使汉语语音合成技术达到真正实用化的水平, 书中给了出四个成功应用的案例. 技术篇还介绍了语音库的制作、合成语音的评测技术, 以及对语音合成技术的展望.

直至今日, 在我国声学界还没有一本体系完整、内容全面的汉语合成专著出版. 该书的四位作者都是长期工作在语音合成的研究开发第一线、专门从事汉语语音声学特征研究和汉语语音合成技术开发的科技工作者, 具有很高的理论水平和丰富的工作经验. 该书对有志从事语音合成研究的青年学者是一个高起点的入门向导, 对学有所成的研究者以及从事相关技术开发的工程师, 也有很好的借鉴意义. 相信该书的出版能够为推动语音合成技术的更大发展、达到人与计算机之间的自然沟通、人人平等享用信息技术, 乃至构建和谐社会的国家目标做出贡献.



中国声学学会理事长

2011 年 5 月 27 日

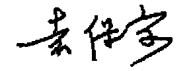
序　　二

20 世纪信息科学与计算机技术的迅速发展, 使得人们对言语工程充满种种憧憬. 让计算机能与人进行言语形式的交流, 成为科学家的研究目标. 语音识别、语音理解与语音合成等项目, 被列入众多国家的研究计划. 其中语音合成技术的成果, 率先走出实验室, 面向各类应用, 开始被公众所认可, 如播报航班、车次、天气、路况信息、各类问询系统, 乃至书报阅读等. 如今, 人们已经可以享受到语音合成技术带来的便利. 尽管技术本身仍有许多值得改进的可能性, 尤其是面对我国汉语发音的丰富多变性与地域环境复杂性的挑战, 但是汉语的合成已经取得了长足的发展. 普及并推广汉语语音合成的有关原理与技术, 已成为当前的燃眉之急. 该书的出版正是迎合了这种技术发展潮流的需求.

该书作者自 20 世纪 80 年代开始从事汉语语音合成的技术研究. 这个团队从汉语发音的声学特征、合成方法、系统研制, 以及产品开发等方面都开展了全面的工作, 取得了一批学术成果, 并在与语言声学界学术交流活动中, 得到了同行的认可; 同时, 也贡献了多款汉语合成与文语转换系统产品, 供各界应用, 取得了一定的业绩.

该书的组织分为基础篇和专题篇两部分：基础篇介绍了声学语音学中关于语音产生、感知、特征分析以及语音学方面的基础内容, 介绍了汉语普通话的韵律特征以及汉语合成系统 (文语转换系统) 的设计实现. 这些内容构成了该书原理性工作的基础, 也融合了不少作者的学术贡献. 专题篇内容着重收录了该团队多年研究的经验及成果, 特别是汉语韵律与数据库建设方面的内容, 有重要的参考价值. 言语工程是一门交叉学科, 需要众多学科的交叉融合. 该书的出版也正是这种交叉的体现, 同时也架起了学科间互为借鉴的桥梁.

愿以该书为契机, 引起读者的兴趣, 有更多的青年学者投入到言语工程, 特别是语音合成的事业中来, 把汉语语音合成推向更高的发展阶段.



2010 年夏于北京交通大学

前　言

“语音合成” 是指一种用机械的或电子的手段, 产生人造语音的技术. 创造一个会说话的机器是人类早就有的梦想. 中国科学院院士马大猷教授说: 古希腊人相信语言是神赐予的, “会说话” 的神像引起不少人的顶礼膜拜; 我国唐朝也有利用木和尚 “说话” 来化缘的记事. 唐朝木和尚说话化缘的故事要比语音合成史料记载的 Kratzenstein 发明元音共鸣腔 (1786 年) 早 1000 多年.

追溯语音合成的历史, 除了中国的木和尚以外, 最早的讲话机诞生于 1791 年, 由 von Kempelen 发明. 随后, Dudley 于 1939 年发明了电子式语音合成器 VODER. 这两件作品向世人证明了语音是可以合成的. 虽然随后有许多模拟式的语音合成器问世, 但都是实验室产品, 要达到实用化, 还必须提高合成速度和合成语音的音质. 在客观上, 20 世纪 80 年代开始的计算机的运算速度和存储能力的显著提高, 为实用化语音合成的发展提供了先决条件. 在提高合成语音的清晰度和自然度方面, 五十多年来各国科学家为此作了不懈的努力. 其中 Klatt 博士创建的串/并联共振峰合成器和日本 ATR 提出的不等长真人语音片段拼接合成技术最受关注. 目前, 在 PC 机上可以做到上百个通道平行实时合成, 语音的质量有一部分可以达到 “以假乱真” 的程度, 大部分也都能达到可被听众接受的水平, 合成语音已开始成为信息交流的一种手段, 成为大众传媒的一分子, 使人类的又一个千年梦想得以实现.

让机器也能说话, 不仅仅赋予机器神秘的色彩, 实际上是人类文明发展的需求和必然的趋势. 众所周知, 21 世纪进入了信息社会, 大量的信息铺天盖地地向我们涌来. 人类生活, 由于有丰富多彩的信息的滋养, 更加充满乐趣, 信息已与空气和水一样成为现代社会生活的必需品. 科学家呼吁 “人人平等地享用信息”. 可是目前信息的传递绝大部分是以文字为介质, 由于社会成员中, 公民的体质和文化素质不同, 不能做到人人平等地享用信息, 这是一个文明社会不允许的. 所以要开发一种技术, 能把文字变成语音, 使得世界上一切用文字传播的信息, 同时也有语音的表现方式. 这种技术装置统称为 “文语转换系统 (text-to-speech system, TTS)”. 要使文字和语音两种信息载体的并存, 让盲人和文盲都能和普通人一样读书、看报. 此外, 信息技术的发展是和计算机分不开的, 计算机推动了人类社会的进步, 提高了人的生活品质. 但直至今日, 计算机还是高技术产品, 被封闭在高知识阶层统治的 “象牙塔” 中, 这同样损伤了部分公民, 而且是绝大部分公民的信息享用权. 要使计算机得到普及, 进入平常百姓家, 必须改变现有的人机界面. 应当使得人和计算机的交际能与人跟人交际一样方便, 即让计算机能听懂人的言语, 同时使计算机也能

说话, 建立人机语音通信接口. 其核心技术是语音识别和语音合成.

21 世纪伊始, 有些激进的科学家提出人类文明的进步将进入一个由 "印刷文化" 向 "口语文化" 回归的新时代. 他们主张, "文字" 仅仅是扩充人脑记忆的一种技术, 而且 "语音" 是天赋的本能. 技术是不断进步、不断更新的, 而本能是和物种共存、不会消失的. 从人类社会的思想表达和交际来看, 语音在先, 文字在后. 但作为纪录语音的文字在历史上有过重要的贡献, 它让我们知道祖先们曾经想过什么, 做过什么? 在无线通信普及以前的漫长岁月中, 让我们能与处在听不到声音的地方的人们进行信息交流, 依靠的是文字. 但是直至今日, 文字也远没有达到尽善尽美的程度, 当今文字所纪录的只是部分的、有限的语音信息, 文字体系的不完善, 阻碍了信息技术进一步发展. 一些科学家提出创造一种新的、和语音一致的、适合计算机的书写体系. 另一些科学家则认为, 回归到口语文化是一个最终的合理解决, 因为语音交际比文字更容易、更有效, 而且不需要专门的学习. 他们举出一些例子证明人类社会正在向 "口语文化" 回归, 比如现代社会中青年人对写作技能训练的厌倦心理和阅读兴趣的减少. 他们预测 2050 年前, 在发达国家具有 VIVO(voice input and voice output) 功能的计算机将进入人类的主流社会, 通过 VIVO 人人都会获得平等地享用信息的权利. 在这种思想的指导下, 2004 年欧盟开始实施一项名为 SAFIR(speech automatic friendly interface realisations) 的计划, 其目标是实现每一个欧盟成员国公民随时随地通过简单的设备, 如电视机、电话机、手机和掌上机等, 用自己的口语与政府的数据网沟通. 每个人, 不管你是有知识的、掌握计算机的, 还是不识字的、对计算机一窍不通的, 都能与网络计算机对话, 自动获取政府网的信息 (包括政府新的公告和法规等), 自动将信息发送给政府相关部门 (如申请书、车辆登记表和所得税税单等). 这项技术还特别适用于某些特殊人群, 如流动中的警察、消防队员和卫生防疫员等, 他们通过简单的移动通信设备可以随时随地获得政府网上的信息, 用于及时正确地处理事故, 发送新信息, 及时更新网上的数据. 欧盟一些地区的实验结果表明, 这项技术显著地提高了政府的办事效率, 节省了开支. 2005 年开始, 中国与欧盟的合作也参加到这项技术的研究中. 研究和开发语音合成技术的目标是实现无障碍的人机通信.

我们从 20 世纪 80 年代介入语音合成技术, 当时并没有意识到语音合成技术对人类文明发展有如此重大的影响. 我们的研究在合成技术方面还是借鉴当时国外的成熟技术, 我们分别利用 HOLMES 和 KLATT 合成器合成了汉语普通话. 也利用真人语音拼接技术和通过 PSOLA 韵律调节合成的普通话, 显著地提高了合成语音质量, 使这项技术达到初步实用化的水平, 推出了 "联想佳音" 汉语合成产品. 但我们的关注点还是在于提高与合成语音的质量相关的汉语声学特征的研究方面, 对影响合成语音的自然度的主要因素进行了综合考察, 发现音高和音长等韵律特征是最重要的因素. 此后, 尤其是转入拼接合成后, 对汉语的韵律特征的研究成为我们

研究的焦点. 我们研究了汉语韵律的层级结构、韵律边界的声学表征、汉语语句的音高下倾特征和语句重音的声学表现. 我们特别关注汉语的节律和重音, 用语料库语言学的研究方法对它作了专题研究. 在此基础上提出了软预测的韵律控制策略, 和韵律导向的选音策略. 在 TTS 系统中实施了如决策树、神经网络、HMM 等基于数据驱动的机器自学习算法, 使得所开发的 "木兰"TTS 系统成为汉语合成技术发展史上具有里程碑意义的经典之作.

在汉语 TTS 中, 语音数据库是基础, 尤其是拼接合成和数据驱动的韵律控制方法. 我们在发音人的选择, 语音样品的采集、标注、编辑和储存等方面积累了一定的经验, 目前几个重要的汉语 TTS 系统的语音库都是我们制作、提供的; 建立了语音数据采集团队和海内外数据采集站点, 为全世界提供多语种语音数据采集服务; 同时我们也十分关心合成语音质量的评测, 对主观的和客观的评测方法做了研究.

这些研究工作的结果发表后得到同行学者的鼓励和关注. 我们在研发汉语合成的道路上走了一程, 有艰辛和痛苦, 也有欢乐, 觉得有责任把我们的研究介绍给读者, 与读者共享. 本书以作者 20 年来语音合成研究积累的经验为基础, 介绍语音合成的原理和方法, 并对文语转换系统中的关键技术进行深入的专题论述. 一方面, 本书具有 "导论" 的性质, 以满足对语音合成技术有兴趣的教师、学生、技术人员、企业家、管理人员等扩大知识面的要求; 另一方面, 本书还有一定前瞻性的专题论述, 为专业研究和工程技术人员提供参考, 以期这个领域的研究和开发能建立在一个较高的起点上, 推动我国科学和技术的高速发展.

技术的发展日新月异, 不能做到面面俱到, 由于我们知识水平有限, 书中难免有不足之处, 敬请读者批评指正.

吕士楠

2011 年 10 月

目　录

基　础　篇

专　题　篇

基　礎　篇

第 1 章　语音合成技术史的叙述

语音合成技术的发展可以作为现代科技发展史的一个范例. 2004 年在匈牙利布达佩斯召开国际会议, 纪念“讲话机”的发明人 Wolfgang von Kempelen. 与会者热烈赞扬他的聪明才智和光辉业绩. 西方人把发明讲话机作为语音合成技术的起点. 今天, 语音合成技术有了长足进步, 合成语音的品质越来越高, 可以达到以假乱真的程度. 它的应用越来越广泛, 在日常生活中听到汽车、照相机在说话, 也不足为奇. 自驾车出门旅行, 卫星导航系统里传出的清晰的语音、热情的服务省却了多少查地图和下车问询的麻烦. 人们开始享用合成语音给我们的生活带来的便捷, 难怪人们欢呼雀跃, 高喊“人类百年梦想实现了!”

其实, 萌发让机器也能说话的思想要比发明讲话机早得多. 我国唐朝就有木和尚说话化缘的记事 (马大猷, 1984), 根据田时秀的考证 (田时秀, 1976), 记载在唐书《朝野佥载》上, (作者张鹭, 唐代文学家 (660~740 年), 调露进士, 官司门员外郎. 开元中, 流放南岭. 撰有《朝野佥载》、《龙筋凤髓》等书.) 故事很简短, 转录如下:

将作大匠杨务廉甚有巧思, 常於沁州市内刻木作僧, 手执一碗, 自能行乞. 碗中钱满, 关键忽发, 自然作声云“布施”. 市人竞观, 欲其作声, 施者日盈数千矣.

故事的人物、地点确切, 叙事明白合理. 至于时间, 按作者的出生和逝世年份推算, 在公元 700 年前后, 比 Kempelen 的讲话机 (1791 年) 至少早 1000 年. 所以说, 今天语音合成技术的发展, “实现了人类的千年梦想”也不为过.

但有实物、有技术文档可考的语音合成发展史, 只能追溯到 18 世纪 70 年代. 语音合成经历了机械式、电子式和计算机语音合成三个阶段, 历时两个多世纪, 融合着几十代人的辛劳, 是一段非常值得回顾的历史.

1.1　机械式语音合成器

1.1.1　Kempelen 的讲话机

在欧洲, 用机器产生语音的研究始于 18 世纪后半叶. 1773 年哥本哈根的生理学教授 C. G. Kratzenstein, 用连接在管风琴上的共鸣管生成元音, 荣获圣彼德堡皇家科学院的奖励. 此时 Wolfgang von Kempelen 已经有了构建讲话机器的意图. Kempelen 是一名在维也纳为 Maria Theresa 皇后做事的工匠. 他于 1734 年生于布拉迪斯拉法, 即以后匈牙利的首都, 1804 年死于维也纳. 他以多才多艺闻名于世,