计 算 机 语 音 技 术

（修订版）

朱民雄　闻　新　黄健群　周　露　编著

北京航空航天大学出版社
http://www.buaapress.com.cn

内 容 简 介

本书全面系统地阐述语音技术的基础、原理、方法和应用。分为3大部分：语音技术的历史、发展和应用的概况；语音技术的生理学、语音学和汉语语音的基础知识及语音过程的声学模型、数学模型和电模型；语音技术中的分析技术、存储与再生技术、合成技术和识别技术。修订版中更新和补充了许多内容，如人-机语音通讯的最新学术动态，新型的第4代语图仪、新型存储器、ISD器件及其应用，可视语言的语音合成技术，计算机语音增强技术及计算机语音技术的新理论和新方法。

特点是：内容系统、叙述清楚、实用性强、有学术研究的参考价值。

本书可供从事人工智能、模式识别、信息与控制、计算机应用的科技人员阅读，也可供高等院校有关专业的教师、研究生及高年级学生参考。

图书在版编目(CIP)数据

计算机语音技术/朱民雄等编著. —2版(修订版)
北京：北京航空航天大学出版社，2002.1
ISBN 7-81077-129-9

Ⅰ.计… Ⅱ.朱… Ⅲ.语音数据处理
Ⅳ.TN912.3

中国版本图书馆CIP数据核字(2001)第074681号

计算机语音技术(修订版)
朱民雄 闻 新 黄健群 周 露 编著
责任编辑 胡 敏

*

北京航空航天大学出版社出版发行
北京市海淀区学院路37号(100083) 发行部电话：(010)82317024 发行部传真：(010)82328026
http://www.buaapress.com.cn
E-mail:pressell@publica.bj.cninfo.net
河北省涿州市新华印刷厂印装 各地书店经销
开本：787 mm×1 092 mm 1/16 印张：24.25 字数：621千字
2002年1月第1版 2002年1月第1次印刷 印数：5 000册
ISBN 7-81077-129-9/TN·003 定价：39.00元

修订版前言

自本书初版本出版至今已近十年。本书发行到全国各地，经三四年时间，已告售罄。但仍有不少读者对本书内容很感兴趣，可惜已很难买到本书了。

岁月在飞逝，时代在前进。在新世纪到来之际，作者经过精心准备，出版本书的修订版。一方面是因为十多年来科学技术飞速发展，涉及本书的各学科领域的研究取得了许多最新成果，作者应该更新和补充内容来进一步完善本书。另一方面是作者应该报答读者们对本书的厚爱，将经过更新和补充的修订版，尽早奉献给读者们。

本书的作者集体扩大到 4 人，新增加了数学、信息技术、测控技术、航天技术等方面的博士、博士后、副教授等高级专业人才，以适应本书涉及的广泛学科领域的需要，为修订好本书奠定了良好的基础。

与本书的初版本相比，修订版中更新和补充了下述内容：

1. 在第一章概述中增加了 20 世纪 90 年代以来人机语音通讯的学术动态，补充了有关的最新研究成果。

2. 在第四章计算机语音分析技术的 4.4 语谱图小节中，补充了新型的第 4 代语图仪 DSP5500 的内容。

3. 在第五章计算机语音存储与再生技术的 5.2 语音信号的存储技术小节中，补充了近年来的新型存储器；5.3 计算机语音处理机小节中补充了语音芯片分类表；新增加了 5.4 新型器件及其应用小节。

4. 在第六章计算机语音合成技术中新增加了 6.4 可视语言的语音合成技术小节。

5. 新增加了第八章计算机语音增强技术一章。

6. 新增加了第九章计算机语音技术的新理论和新方法一章。

7. 补充修订了本书末的参考文献。

本书初版本出版后，许多读者通过北京航空航天大学出版社，用书信、电话或面见等方式与作者取得了联系，咨询或讨论本书中的有关内容，作者对此深表感谢。作者赞赏读者的渴求和探索科学知识的精神。作者希望，在本书修订版出版后，能继续得到广大读者的欢迎和支持。

作者竭诚欢迎相关专家和学者们对本书内容的批评和指正。作者愿与广大读者一起去探索新科学技术知识，推动计算机语音技术的更大发展，振兴中华，将我国建设成世界一流的强国。

朱民雄

2001 年 7 月于北京航空航天大学

目　　录

第一章　概　述

自从 20 世纪 70 年代第 1 块微处理器芯片诞生以来，微型计算机技术日益发展，已经渗入到许多领域，得到了广泛的应用。微机技术渗入到声学领域，它与语言声学相结合，使语音通信进入了发展的新阶段。语言是人类相互间进行通信的最自然和最方便的形式，语音通信是一种理想的人机通信方式。语音通讯的研究涉及到人工智能、模式识别、数字信号处理、微机技术、语言声学、语言学和认知科学等许多学科领域，是 1 个多学科的综合性研究领域，其研究成果具有重要的学术价值和应用价值。

计算机语音技术是语音通信领域的 1 个重要部分，包括 4 种技术，即语音分析技术、语音存储与再生技术、语音合成技术和语音识别技术。从语音通信涉及的内容而言，语音技术还应包括语音理解技术。但是，在学术上，由于历史的原因，长期以来，语音识别和自然语言处理（包括语音理解）两个研究领域是并行独立发展的。目前，主要的研究工作还是语音识别。虽然，美国 DARPA 战略计算计划提出了研究口语系统（Spoken Language System），该系统要求把语音识别和自然语言理解结合起来，并进一步实用化。但这是下一代语音识别系统。基于上述情况，本书仅涉及传统的 4 种语音技术。

现简单介绍一下语音技术的发展概况。对于人类语音发生过程的研究可以追溯到很早的年代。那时，人们研究人类发声的物理过程及其数学表达方式和模型。另一方面，人们还研究语言语音学，了解语音的分类、性质、表示方式等。20 世纪 30 年代到 40 年代，美国 Bell 电话实验室的研究人员在主任 O. E. Buckley 的支持下，对英语语音分析作了大量的研究工作，取得了一些重要成果。其中有些成果对我们当前的工作，仍有相当重要的指导意义。语音技术最早和最重要的一种应用是 Homer Dudley 在 1930 年发明的声码器。他在 1939 年以“Remaking Speech”为题的论文发表了这个成果。1949 年贝尔实验室的研究人员研制成功第 1 个电合成器。他们把它叫作电发声系统（EVT），这是把有限的双管谐振模型（双亥尔姆霍兹谐振模型 Double Helmholtz）的声学特征转换成为电气等效电路。它只能发英语元音，但是，用实验证实了在一定条件下，双管谐振模型是正确的。对语音识别的研究，可以追溯到 50 年代。1952 年 Davis 等人研究成功了世界上第 1 个识别 10 个英文数字发音的实验系统。1960 年 Denes 等人研究成功了第 1 个计算机语音识别系统。

进入 70 年代以后，语音技术取得了许多实质性的进展：① 用于语音信号的信息压缩和特征提取的线性预测分析技术；② 用于以线性预测编码表示语音参数时相似度测量的线性预测残差；③ 用于输入语音与参考样本之间时间匹配的动态规划方法；④ 一种新的基于聚类分析的数据压缩编码的矢量量化方法等。

在 70 年代，语音技术的产品首次进入商品市场。1976 年 Votrax 推出 Computalker 语音合成器进入计算机业余爱好者市场。它采用 8080 微处理器，并用 S－100 总线与其他许多微计算机系统连接。它有 6 KB 的存储器存储音素表和程序，以机器可读的标准语音表的代码输入。虽然，Computalker 产生的合成语音质量很差，但是，合成语音已被广大个人计算机用户所接受。1976 年 Votrax 公司推出另外 1 个产品 ML－1 语音合成器。它的早期产品为

VS－6型。它们都是规则合成语音的最早产品。ML－Ⅰ型采用80个音节、8级音高和4级不同发音持续时间。ML－Ⅰ型使用手册还给出了一份625个单词和短语的章节词典。在80年代初，Votrax公司推出大规模集成电路芯片SC－01型，它采用音素合成技术。1978年夏，TI公司首次推出单片语音合成器，型号为TMC0280，它采用超大规模集成电路技术。这一产品使TI公司遥遥领先于它的同行，并使语音领域的许多专家惊奇不已。TI公司用此芯片推出了一种产品，叫Speak'n Spell toy。这种售价为50美金的产品，使语音技术走出研究实验室进入消费者市场。此产品的面板上有26个字母键和14个附加控制键。它采用4位微处理器TMS1000，2个128 kbit的的ROM内存约330个单词和短语(语音持续约3～4 min)，数据传输率为1 200位/秒。它采用线性预测合成方法，由格型滤波器实现，格型滤波器有10级用10个反射系数表示。语音合成的控制参数有12个：10个反射系数，1个音高参数和1个能量参数。在1978年TI公司的会讲话的Speak'n Spell toy出现以后，又有许多基于微机的会讲话的产品推出。如会讲话的怀表，会讲话的微波灶，会讲话的弹球机，会讲话的计算器等。

进入80年代，国外对语音技术的研究和开发更加活跃。大学和研究所一般致力于学科前沿的研究，而大公司则着眼于市场需要，致力于开发实用化的商品。在语音识别技术方面，小词汇量特定人孤立词语音识别技术已经发展成熟。每年生产数以百计的语音识别商品，用于工业、军事以及医疗部门的指挥岗位、产品检验岗位、数据录入岗位及其他一些手眼并用的场合，用作口呼命令、口呼数据录入以及向计算机或其他机器传递信息。同时，还研究成功了大词汇量、非特定人、连续语音识别实验系统。如美国Dragon公司研制成功Dragon Dictate系统(已投放市场)，它的技术性能如下：

词汇量	30 000个词，可扩充
说话人训练方式	说话人自适应
说话方式	孤立词
识别方法	概率模型
语法限制	自然语法
识别率	90％(20 000个词)
实时性	40个词/分
硬件要求	AT/386机，6 MB内存

它具有下列特点：

- 与PC相兼容。
- 开放式词汇表，用户可以方便地调整和扩充新词。
- 具有说话人自适应能力，新用户不需对全部词汇表进行训练，在使用中不断提高识别率。
- 具有很强的人机交互能力或友好用户接口。

另外1个成功的系统是美国卡内基–梅隆大学的Sphinx系统，它能识别997个词汇，在非特定人的条件下，识别率可以达到94％。它的特点是：

- 集成前人的研究成果，实现系统各个环节的优化。
- 应用多种知识源(声学、音素、词汇、句法、词义)，提高系统的区分能力。
- 语音信号中提取多种特征，提高系统的识别率。
- 引入反映协同效应的音素模型，减少语音信号多变性的影响。

国外的大公司已开发了性能较好的产品投入商品市场。Bell 实验室的 Conversant 语音信息系统就是这类有代表性的产品之一。它的性能如下：

词汇量 0～9(包括 o 的两个读法：zero 和 oh)，yes，no 共 13 个词。

特 点 词汇量极少，识别率高，鲁棒性高。非特定人，数码可流利地连续。具有过滤背景噪音及非语音信息能力。关键词跟踪。

用 途 电话定货，股票交易，查询银行帐目等。

另一个有代表性的产品是美国 TI 公司的 TM 英语博士(English Professor)。它的外形与该公司早期产品 Speak'n Spell toy 相同，其面板上有 26 个英语字母键和 14 个控制键。用液晶显示英语，并有喇叭输出英语和汉语普通话语音。它有 3 种工作方式：读音——按英语字母键，输入英语某单词的拼法，在液晶显示器上有该字的文字显示，按“输入”键，可听到该单词的英语读音和汉语释义词读音。听音选字——在喇叭中听到某英语单词的读音后，在液晶显示器上显示多种英语单词，按“输入”键选出正确答案，而后有该词的汉语释义词读音输出。拼字——听到某英语单词的读音后，输入该词的拼法，而后有该词的汉语释义词读音输出。它有 3 个不同的词库模块，可按学习需要更换模块。

今后，语音技术的研究将更为深入。下一代语音识别系统要求把语音识别和自然语言理解结合起来。例如，美国 DARPA 的战略计算计划中的口语系统正在进行这方面的研究。到 1993 年它要达到的具体要求是：能理解对话型自然语音(口语)；词汇量为 5 000 个词；困惑度为 100～200；语音处理与自然语音处理完全结合；实时；任务完成率 85%；多应用领域，有可修改的用户接口；对各种因素引起的语音信号变动的鲁棒性。

上述系统的实用系统也正在研制。目前，Bell 实验室和 MIT 正在按上述要求研制民航定票信息系统 ATIS(Air Ticket Information System)。

在我国，语音技术的研究起步较晚，投入的研究单位和人员也比较少。目前，我国开展这方面研究工作的人员，正在跟踪先进国家在这一研究领域的最新动态，努力赶上世界先进水平。

在我国，语音技术的产品较少，技术性能也比较差，功能较简单，应用领域也比较少。我国的语音技术的产品分为两大类：语音合成技术的产品和语音识别技术的产品。语音合成技术的产品有下列几类：

1. 数字语音留言机

采用语音信号存储与再生技术。它属于时域波形编码的语音合成，编码方法采用 ADM 和 ADPCM 两种。采样频率为 4～8 kHz，最大可达 32 kHz。存储时间有 8 s、16 s、40 s、128 s 等。存储器用 1～4 个 256 kbit 的 DRAM。5～6 V 直流供电。由于存储器的容量小，存储时间短。这种产品的应用范围较小。今后，在大容量存储器，如 4 Mbit 以上容量的价格比较便宜，则其应用会更广泛，如用于电话留言机，甚至替代目前的卡式磁带录音机。

2. 电脑报站机

这种产品已比较广泛地应用于公共交通汽车、地铁列车等上。语音信号波形经过压缩处理变成数码存储于存储器中。使用时由按键给出指令，在控制软件的管理下，根据指令需要，把数码合成为语音信号输出。这种产品功能齐全，工作可靠，使用方便。

3. 电脑语音报警器

在冶金、化工、石油、电力等工业的自动控制系统中，在各种仪器仪表中，在机器人中都需

要有报警信号输出。过去,常用的是声响或闪光报警信号。新一代报警器是电脑语音报警器,它具有报警意义明确、工作可靠、可远距离传送、使用方便等优点。汽车用电脑语音报警器就是一种具有上述优点的报警器。它采集汽车中常用的工作参数(如冷却水温度、汽油存储量等)信号及各种灯光(如刹车灯、近光灯、左转灯、右转灯、长明灯等)信号,一旦这些信号到达预定的报警数值,则在喇叭中输出相应的标准汉语语音信号,以向司机提示。这种报警可以延长汽车使用寿命,遵守交通安全法规,减轻司机在开车时的精神和心理负担。

4. 语音合成卡

本产品采用线性预测编码技术压缩语音数据,用声韵母音元拼接合成全部汉语音节,包括了国标 GB2312-80 的一、二级汉字发音。它插在 PC 机及其它兼容机内,由专用合成软件,在键盘输入或屏幕显示的同时输出汉语语音。它可应用于需要计算机进行语音输出的场合,如计算机辅助教学、文稿校对、自动化系统检测和报警等。另外一种文本阅读系统也是这一类的产品。

语音识别技术的产品也有几种。中西文语音识别系统是一种在 PC 机及各种兼容机上使用的产品。它能将人类的语音自动转化为文字或指令,用于快速录入汉字、声控操作计算机或电子打印机。其主要性能指标如下:

容量　　一、二级汉字和 2 000 条口令

识别率　　词组为 98%,单词为 90%

识别速度　每分钟 80 字左右

抗噪能力　不高于 75 dB(A)

这种系统为特定人语音识别,识别率高、实时性好。但抗噪声能力较差,在高噪声下,识别率会有较大下降。另外使用时要有 PC 机或其他兼容机,整个系统价格就较高。因此,它的应用场合受到限制。

在本书首次出版时,已经召开了第一届全国人机语音通讯学术会议(以后为每两年举行 1 次)。全国人机语言通讯学术会议是由中国自动化学会模式识别和机器智能专业委员会、中国计算机学会人工智能与模式识别专业委员会、中国电子学会信号处理学会语音图像通讯专业委员会、中国声学学会语言听觉和音乐声学分科学会、中国中文信息学会基础理论专业委员会、中国通信学会通信理论专业委员会和国家高技术智能计算机系统专家组联合主办的全国性学术会议。它是全国人机语音通信研究领域最为集中和最为重要的会议之一,受到全国有关同行的重视和关注。会议取得了很大成功,促进了国内有关领域的学术交流,扩大了本专业在国内外科技界的影响。最近的两届会议是 1996 年召开的第四届全国人机语音通讯学术会议(NCMMSC-96)和 1998 年召开的第五届全国人机语音通讯学术会议(NCMMSC-98)。第四届会议有更多台湾、香港以及国外的访问学者的积极参与,也受到海外华人、专家学者的关注。第四届会议入选论文 73 篇,其中属于语音分析有 8 篇论文,属于语音存储的有 11 篇论文,属于语音合成的有 5 篇论文,属于语音识别的有 42 篇论文,还有其他相关论文 7 篇。这些论文在相当程度上反映了我国人机语音通讯学术界在理论和技术两方面的最新进展,也是我国语音信号处理领域成果交流的集中体现,表明我国人机语音通讯科研工作在攀登高峰的前进路上又向前迈进了一步。第五届会议的主要内容是交流全国在语音识别、合成及信号处理等方面的学术研究与开发成果,研讨人机语音通讯技术的发展方向和实用化途径。第五届会议入选论文 82 篇,其中属于语音分析的有 19 篇论文,属于语音存储的有 12 篇论文,属于语音

合成的有 11 篇论文,属于语音识别的有 34 篇论文。这些论文反映了我国近年来人机语音通讯技术方面的研究开发工作和最新成果,对我国人机语音通讯技术的发展有重要参考价值,进一步推动我国的人机语音通讯的研究和实用化进程。历届会议的召开都取得了很大成功,对我国人机语音通讯的研究有重要意义。

根据本书写作的宗旨和大多数读者特别关心的内容,本书将介绍最近两届全国人机语音通讯学术会议的部分论文内容。

1. VASTP——缝纫机声控装置

由美国、中国、日本 3 国的 3 位学者撰文。VASTP(Voice Asisted Sewing Technology Product)是语音识别技术在服装工业中的具体应用。它接受操作人员的语音指令,用来控制缝纫机的有关操作,从而解放了操作人员的双脚,实现了现代服装生产流水线工人的离机操作,减轻了工作强度,并为残疾人提供了在服装制造业就业的机会。该装置所采用的相关语音技术如下:

(1) 采用与说话人有关的孤立词模式匹配法语音识别技术

语音输入经一个滤波器组(ASA16 芯片)获取短时谱语音信号,滤波器组将频率在 200～7 000 Hz 范围分为 16 个通道滤波器,每个滤波器由 1 个二阶 Butterworth 带通滤波器、1 个半波校正器和 1 个衰减频率为 25 Hz 的二阶低通滤波器组成。由 1 个 16 通道采保多路选择器,以采用间隔为 6.25 ms 读取 16 个语音谱特征参数。识别词汇可达 1 000 字,如"slow"、"medium"、"fast"、"trim"、"raise"、"stop"等。

(2) 采用噪声压缩及自适应软件处理技术

由于工作现场的噪声大,且其变化幅度也大因而需要用噪声压缩及自适应软件处理技术。语音输入经头盔或抗噪声话筒或喉头话筒,输入到 1 个增益可编程的放大器,其增益由 1 个 8 位增益控制的寄存器控制,可作手动、自动实时增益控制,以适应不同灵敏度的话筒和不同噪声的幅度。最后,用自适应技术改善噪声环境下的识别性能。

(3) 采用音素合成(SSI263A 芯片)和 ADPCM 语音波形编码(NEC 7759 芯片)两种语音合成技术

前者是基于音素的语音合成器,通过仔细设计语音合成规则,可直接实现任意英语文本的文语转换。它使用 64 个音素符号(由字母、数字组成),其中 34 个用于美式英语的基本音,27 个用于表示基本音的变音,还有 3 个表示无声状态。芯片为每个音素提供了 8 个运行时可编程的参数。为了得到 1 个词组或短语的语音,用不同音素及其可编程参数以控制音变、幅度及音长等。后者用 ADPCM 编码,在 PROM 的不同组及组号、序号存入单词或短语语音。由该装置的 CPU 来选择组号及序号,重放相应内容。后者比前者的语音音质好,但存储量有限,并需经前期处理,使用不如前者灵活。

2. 傻瓜式声控电话机

由清华大学和五邑大学联合研制。傻瓜式语音拨号 VDD(Voice Diallers for Dummies)是语音识别技术在电话机中的应用。该机的电子电话簿中存有私人电话用户对应姓名的电话号码。打电话时只需口呼对方的姓名,在语音和液晶两种提示下,简便操作就能完成。这种可称为傻瓜机的声控电话机的错误率小于 1%。据报道,20 世纪 80 年代瑞典 Ericsson 公司已装备这种电话机,应用于该公司内部的快呼通讯网。这种电话无论白天黑夜,盲人还是正常人,均可方便使用。该机所采用的语音技术如下:

电话线路的频带很窄(约 4 000 Hz),用 8 kHz 频率对输入语音信号采样,得到 10 阶的 LPC—CEP 系数作为语音识别的基本参数。语音识别模型是连续距离密度的分段概率模型(CDD—SPM)。这种模型占用的存储量小,识别响应时间快。据研制者进行的满容量测试结果表明:直接拨出的正确率为 41.3%,用语音输出将几个侯选姓名顺次报出后,再确认拨出电话号码的正确率为 54.7%,错误率为 1%(错误的情况包括没有经过确认直接拨错的情况和因噪声而被识别成非所要姓名的情况等)。

3. 北京旅游信息咨询系统 VDTIRS

VDTIRS(Voice-Driven Tourism Information Retrieual System)由中科院自动化研究所研制。这是中国科学院"八五"重大科研项目"汉语人—机语音对话系统工程"的一部分。这是 1 个非特定人连续语音人机对话识别系统。有 1 004 个词的词库和 61 个句型的语法限制,用户可以用连续的语音向系统咨询有关北京的名胜古迹、博物馆等场所的信息(包括位置、图片、文字介绍、开放时间、乘车路线及天气等)。系统具有一定的理解和对话功能,并在接近实时条件下,以语音、图像、图形、文字等多媒体形式作出回答和响应。对训练集外的非特定人测试者的现场测试的系统识别率为 94%。该系统所采用的相关语音技术如下:

采用离散 HMM 算法、有限状态文法和 Viterbi Beam 搜索为核心技术。系统采用基于 Mel 尺度的倒谱系数 MFCC 和能量作为语音信号的声学特征,并引入描述信号动态特性的一阶及二阶导数。这些特征通过 4 个码本的矢量量化量(VQ)得到四维的语音特征矢量序列,送到识别器识别。语音识别中的语言模型采用有限状态语法作为系统的语言模型。利用用户提出的任务中的目标和属性和与语法项相关的语义信息,构成有限状态语法网络(也是 1 个 HMM 网络)。系统在声学模型和语言模型处于统一的 HMM 框架下,由语言模型、词的拼音模型和音素的声学模型构成状态图,采用 Viterbi Beam 搜索算法、从状态图中找到 1 条最佳路径作为识别结果。

4. 铁路订票系统

由台湾工业技术研究院电脑与通讯研究所研制。订票者通过电话输入所订票的资料:订票者身份证号码、车种(目前仅限 3 种)、车次、起始终点站名、乘车日期(5～7 日后)、车票张数(最多 4 张)。当所有资料输入完毕后,系统会重播所输入的资料供订票者确认。若已确定系统辨识的错误,则可再输入"修正"一词语音,系统会回到上个提示,让订票者重新输入。若所有确认无误后,则系统会提供 1 个号码,订票者需用此号码到乘车处取票。该系统所采用的相关语音技术如下:

在台湾地区共用 800 个语音建立 MAT(Mandarin Across Taiwan)语音资料库。用 1～4 字词训练模型供系统辨认车种、站名、日期、张数,用连续数字训练模型供系统辨认车次。辨识模型为用大量电话语音训练而得的声韵母连续马尔柯夫模型。共建了 1 个静音、21 个声母、36 个韵母的独立文语模型。系统识别词不达意汇量为 35 个词组(车种与站名),98 个 1～4 位数字(车次),3 个日期及 4 个张数,属于中等词汇量。8 kHz 采样频率。试验表明其辨识率及响应速度尚能接受,但噪声仍是问题。因为电话话筒能接收讯号的范围相当大,稍有干扰即能影响辨识效果。

5. BJD—97 汉语文语转换系统

由北方交通大学研制。此系统利用波形编码合成技术,其合成的数码率大,存储量也大;但其合成的自然度好,且系统结构简单。随着 PC 机的迅猛发展,处理能力和存储功能比以往

有了很大增强。因此,系统是 1 个基于 PC 和 Windows 95 环境下的汉语文语转换系统。该系统所采用的相关语音技术如下:

以"现代汉语词表"和"现代汉语词典"为基础选择了 1 个近十万余条的汉语分词词典。该分词词典最长的词长度为 7 个字,可以覆盖绝大多数的汉语词组。根据分词规则,对语音输入文本进行语言学处理,给出发音描述。采用 64 kb/s 采样频率的 PCM 波形编码建造 1 252 个有调基本四声音节和 378 个基本轻声音节的单音节库和根据"语文学习词典"建造 2 字、3 字、4 字等近一万条的词库,单音节库和词库构成音库。在连续汉语发音时,存在声调上的变化和单音节之间连接的不同。考虑重音的变化,计算出调域、调基值、时长等调整为新的音节基频曲线,计算出每帧波形数据送入缓冲区。合成时从缓冲区中取出送至声霸卡数/模转换后输出语音,直至文本结束。据研制者的测试,工作情况令人满意。该系统由于韵律调整的模式比较单一,输出语音单调,短语之间的直接拼接,合成语音有些机械。这些问题有待进一步解决。

6. 汉英口语机器翻译原形系统

由哈尔滨工业大学研制。口语翻译与文本翻译不同,具有口语不连贯性、语法约束相对较弱、没有明确的句子边界、识别错误率大、附有不同可信度输出的选择等困难和口语对话一般结构不很复杂、句子较短等特点。口语语音翻译系统要实现大词表非特定人连续语音识别、恰当表达讲话者意图的机器翻译、流畅的目标语语音合成输出等,其研究涉及自动化、人工智能、认知科学、思维科学等学科领域。多个国家、多种学科的专家和多种资助渠道等是其研究的特点。1993 年成立的国际语音翻译联合会 C—STAR(Consortium for Speech Translation Advanced Research)是一个以口语语音机器翻译为基本研究目标的国际合作组织,现在共有来自 12 个国家的 20 余个成员参加。该系统所采用的相关语音技术如下:

该翻译系统包括了 3 个模块。首先是基于规则的汉语口语预处理模块。它是根据汉译英词组对应而对汉语音词组进行预处理的分词,并对讲话者口语失语的错误进行基于规则的修复处理。其次是采用精简循环网络的神经网络方法的汉语口语分析模块。它进行词汇级的语法、语义的选择,短语边界的界定和短语级语法、语义的确定。其共有词汇级语法类 16 个、语义类 22 个,短语级语法类 9 个、语义类 21 个。最后是基于实例的汉英口语翻译模块。它将输入语句与从语料库中检索到的合适例句进行相似度(词的相仿度、上下文环境相似度及全句的相似度)计算,选出 1 个相似度最大的例句,参照此实例构造输入句的译文。研制者已完成了 1 个能初步运行的面向会面安排领域的汉英口语语音翻译的原形系统,将继续改进以完善该系统。

2000 年首次在北京召开了中文口语语言处理国际会议(ISCSLP 2000)。这是继 1998 年在新加坡成功召开的 ISCSLP 系列会议的第 2 次会议。下两次会议将在台北(2002 年)和香港(2004 年)召开。本次会议共收到中国、台湾和香港地区、新加坡、日本、美国、加拿大和德国8 个国家和地区的 106 篇高质量的论文。最后入选本次会议的论文有 2 篇综述报告,4 篇重点报告和在 9 个小组上宣读的 85 篇论文。论文的内容丰富,涉及语音技术的各个领域,其中属于语音分析(包括声学建模)的有 30 篇论文,属于语音存储(包括语音翻译)的有 5 篇论文,属于语音合成的有 10 篇论文,属于语音识别(包括语音理解)的有 40 篇论文。这次是在中国召开的首次 ISCSLP 会议,原定 2000 年召开的全国人机语音通讯学术会议没有召开。大部分这一领域的研究论文都投到 ISCSLP 2000 会议上去了。

微软中国研究院的研究人员在会议上作了"自然语言处理(NLP)研究最新动向"的综述报

告。报告者提出了他观察到的一些新动向,如多重特征和复合语法、语法研究的词汇化、统计语言建模和词集研究等,虽然前两种方法还将沿着传统语言学研究方向向前发展,但它们会比以往更加精细地探索语言学知识,特别是词汇知识。最后1种方法是所谓的实验研究,它是基于可观察到的论据而不是语言学的直觉。这些研究的任务是更精细地从大词集中自动地或半自动地获得语言学知识,因为它仍然是NLP技术发展的瓶颈。

朗讯科技贝尔实验室的研究人员在会议上作了“下一次革命:从图形到语音用户接口”的另一综述报告。在电话发明后的最后1个世纪中的大部分时间内,人类已不再满足于曾经是主要通讯技术的简单联络方法了。20世纪末因特网的使用蓬勃发展。现在,人们除了讲话以外,还增加了更多信息和内容的要求。现在存取信息的最通用载体还是使用图形用户接口(GUI),用鼠标点击的传统接口。然而,由于无线和移动的因特网设备的需要,如电池电话、个人数据助手,由于键盘和显示器微型化,这种传统的GUI在有效传送存储信息时仍有诸多困难。另外,语音输入和输出设备在移动通讯器中经常是内装的和大量使用的。由于特别是在眼忙和手忙情况下,语音是最自然的接口和通讯工具,在未来设备的新设计多模式用户接口中,语言将是一种主要方式。这就提出一种补充传统GUI的语音用户接口(VUI)的革命性设计。论文作者述评了包括自动语音识别(ASR)、文语转换(TTS)、讲话人辨认(SV)和发声辨认(UV)的核心语音技术,讨论了每项技术的技术能力和限制。为了建立有效的VUI,对话管理和自然语言理解是两个尚未成熟的关键部分,许多研究的问题必须解决而尚未解决。近年来,完成和开展了大量带有全部制约条件的应用和设备。由于应用需要的增加,建立了一些语音入门公司。论文介绍了一些应用实例并讨论了它们成功的原因。最后,论文说明了网页和集成及语音入门电话设备的最新研究,指出了实现VUI新技术的困难。在新千年的未来信息年代,这些研究提供了巨大的商业机遇和对“地球村”社会的好处。

由于本会议论文集的内容丰富,因此,除了其个别论文将在本书中作介绍外,限于时间有限和内容广泛等原因,本书作者只能请有兴趣的读者自己去查阅、研究这本论文集吧!

第二章　语音技术的基础知识

2.1　语音过程生理学基础知识

2.1.1　语音发送过程生理学基础知识

1. 人类发出的语音波形是一种声压波

语音是由图 2-1 所示的人类发音器官的生理运动所产生的。人类的发音器官及其作用分为以下 5 类：① 喉(振动源)；② 肺(能源)；③ 声道——从喉到唇，包括口腔(谐振源)；④ 鼻腔(谐振源)；⑤ 发音器官，包括唇、齿、齿龈、舌、颌和面颊(改变谐振腔的外形)。当产生语音时，例如发“eve”中的/i/音，空气由肺部压入，由嘴唇呼出，从而引起声门的开启和闭合(声带间的开口定义为声门)。开闭的速率取决于声道中空气压力和声带的生理控制。声门的闭合是由两侧声带和假声带互相接近的结果，二者的接近不仅使声门区闭合，且具有双重的活瓣作用。声带振动，是产生声音的基本声源。声带对气流的阻抗能力大小不同，声带抵抗自上而下的气流冲开声门裂的能力，可数倍于抵抗气流自下向上冲开声门区的能力。

2. 声带的振动决定于其质量

质量愈大，每秒振动愈少；反之，质量愈小，声带振动愈快。声带振动频率决定了声音的音高，高音高声为高频声，妇女和小孩属于这一类。高音高声音是声带质量小的缘故，因而每秒振动频率高。男性的声带振动频率范围为 50～250 Hz，女性的范围约近于 500 Hz。由肺部来的气流经声门区输入到声道，并由唇或鼻输出。在声门区内，下声门的空气压力及其随时间的变化决定了压入声道的声门气流的体积速度(亦称声门体积速度波)。这声门体积速度波为输入到声道的声能或激励函数。声门开闭的速度，在声学测量上近似为所观察到声压波周期的倒数。

图 2-1　人的发音器官简图

3. 关于喉的发声机理有两种学说

这两种学说为：① 张力学说，也称肌-弹力学说。它认为从气管内呼出的气流的压力可使声门裂发生节律性的开闭而使声带振动发出声音；② 阵挛学说，也称神经-肌肉学说，认为声

带振动是中枢神经系发出有效的神经冲动，使声带肌肉发生节律性收缩而产生声音，且认为音调的频率即中枢所传下的神经冲动的频率。近年来许多学者认为张力学说虽不全面，但基本符合发声学现实。我们这里也采用张力学说来说明发声的生理过程。

(1) 喉——喉的主要生理过程是声门区开闭和声带的振动。讲话声音由声带振动或没有声带振动来产生。前者产生浊音(有声音)，所有的元音和一些辅音是浊音；后者产生清音(无声音)，一些辅音是清音。因此喉的声带用于产生浊音。

(2) 肺——人类呼吸系统的生理过程提供了使声带运动的必要能量。当人吸入空气，其肺部扩张，胸腔也扩张。空气从肺部呼出使空气经过喉部，这一能源使声带振动。此外，空气可被阻塞而产生某些语音，可以改变阻塞程度，例如语音/p/或/b/要求全部阻塞，而/f/仅要求部分阻塞，这就是摩擦辅音。声强大小取决于空气经过声带时的能量。肺还可响应声音幅值的要求，使受话者听到声强不同的声音，供给的能量愈大，声带移动也愈大，产生的声波幅值也愈大。

(3) 声道——从喉到唇包括口腔为声道。特别是口腔，对全部发音有重大影响，这是由于舌和颌的移动，使口腔外形有很大的变化。声道是 1 个谐振腔，它放大某个频率而衰减其他分量。声带振动频率决定了基频。谐振频率由每一瞬间的声道外形决定，再迭加其基频。讲话时，舌和唇连续运动，使声道常常改变外形和尺寸，随即改变谐振频率，并使谐波改变，这就使讲话成为 1 个连续变化的声波。声道也会产生讲话声波的能谱峰，这种波峰称为共振峰。当发音器官稳定时出现共振峰，结果声道在 3～4 个泛音频率下谐振。在连续讲话时，由于改变了声道的外形和尺寸，故共振峰也发生改变，但其变化速度较低，这是受我们移动舌、唇、颌等的快速程度的限制。假设从喉到唇的典型距离为 17 cm，音速为 340 m/s，则在 500 Hz、1 500 Hz、2 500 Hz 产生谐振。这些共振峰常称为 F_1(其频率 f_1 约 500 Hz)、F_2(其频率 f_2 约 1 780 Hz)和 F_3(其频率 f_3 约 2 500 Hz)。这些共振峰的计算如下：

设男性成人的声道长 $L=17$ cm，音速 $c=340$ m/s，声波在声道 L 中传播，当声道 $L=\frac{1}{4}\lambda_1、\frac{3}{4}\lambda_2、\frac{3}{5}\lambda_3$ 时，声波在唇处达到最大值并辐射出去。$\lambda_1、\lambda_2、\lambda_3$ 分别为第 1 共振峰频率的波长、第 2 共振峰频率的波长和第 3 共振峰频率的波长。

第 1 共振峰 F_1

波长 $\lambda_1=4L=4\times0.17=0.68$ m

频率 $f_1=\frac{c}{\lambda_1}=\frac{340}{0.68}=500$ Hz

第 2 共振峰 F_2

波长 $\lambda_2=\frac{4}{3}L=0.226\ 7$ m

频率 $f_2=\frac{c}{\lambda_2}=1\ 500$ Hz

第 3 共振峰 F_3

波长 $\lambda_3=\frac{5}{3}L=0.136$ m

频率 $f_3=\frac{c}{\lambda_3}=2\ 500$ Hz