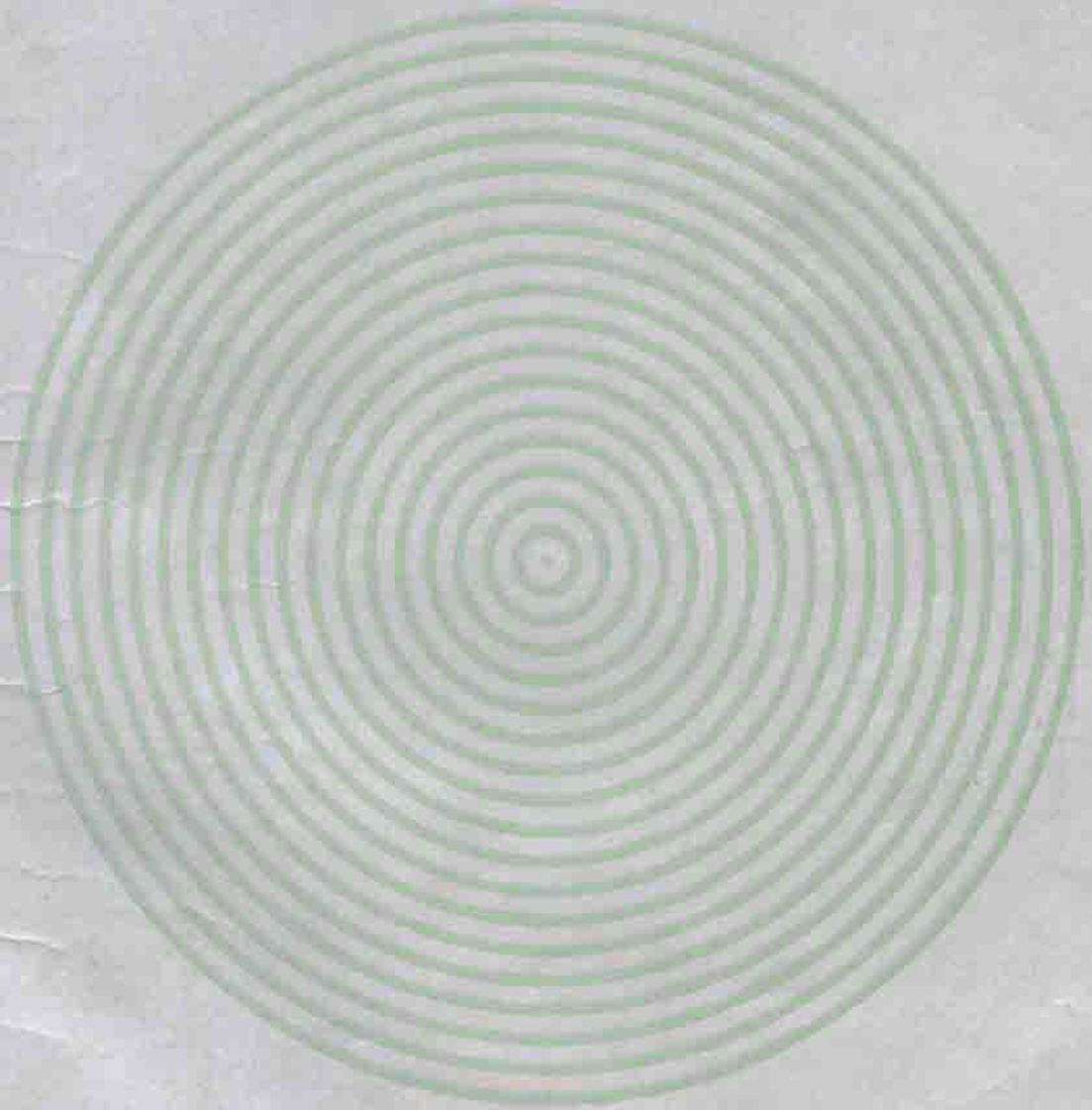
计算机语音技术

朱民雄　编著



北京航空航天大学出版社

计算机语音技术

计算机语音技术

朱民雄 编著

北京航空航天大学出版社

（京）新登字166号

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。分为三大部分：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。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。

计 算 机 语 音 技 术

JISUANJI YUYIN JISHU

编　　著　朱民雄

责任编辑　赵延永

北京航空航天大学出版社出版

新华书店总店科技发行所发行　各地新华书店经销

北京农业工程大学印刷厂印装

787×1092　1/16　印张：14.5　字数：371千字

1992年8月第一版　1992年8月第一次印刷　印数：7000册

ISBN 7-81012-321-1/TP·067　定价：8.70元

前　言

本世纪已进入最后10年。现代科学技术正在迅速发展，它使世界发生了许多惊人的变化，目前的时代是信息的时代。信息的交换——通讯是一个活跃而又热门的研究领域。语言是人类相互间进行通信最自然、最方便的形式。自然语言——语音是一种理想的人机通信方式。它的实现，可以为计算机、机器人和其它自动化系统建立更为良好的人机交互环境，可进一步推动计算机和其它智能机器的应用，提高社会信息化和自动化的程度，将人类的物质生活水平和精神文化水平推向更高的阶段。

语音技术是人机通讯的一种方式。它涉及到人工智能、模式识别、数字信号处理、语言声学、语言学和认知科学等许多学术领域。人机语音通讯的研究受到了学术界的重视。其研究成果转换为市场需要的产品，也受到了企业界的关注。

语音技术的研究已有近半个世纪的历史。语音技术产品的开发也已有20年的历史。这种研究和开发发展很快。我国自执行改革和开放政策以来，科学技术现代化取得了很大的进展。在语音技术的研究和开发方面，也有了很大的发展。关于语言技术方面的专著，在国外已出版了10多本。在我国，除了几本译著外，由我国的专家学者写的有关语音技术的专著甚少。作者编著本书就是为弥补这种不足的一种尝试。

本书是一本全面系统地阐述语音技术的基础、原理、方法和应用的专著。第一部分主要叙述语音技术的历史，发展和应用概况。第二部分由第二章和第三章组成，主要介绍语音技术的生理学、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。第三部分由第四章到第七章组成，顺序地叙述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。

本书的特点概述如下：

1. 具有系统性。语音技术方面分别介绍语音分析、语音存储与再生、语音合成和语音识别四种技术，其中语音分析技术又是后三种语音技术的基础。读者按本书的顺序阅读下来，容易掌握和了解语音技术的来龙去脉，获得比较系统知识。

2. 叙述简单明了，通俗易懂，基本概念和重点突出。对于一些难点，则作了详细的分析，配合以图表、实例，使读者容易掌握这些知识。

3. 具有实用性。本书对各项技术均给出了许多实用知识，如所用集成电路芯片，典型应用电路，电路的设计和分析方法，实验数据及结果分析等。

4. 作者将多年来在语音技术方面的一些研究成果在本书中介绍出来。其中的绝大部分是首次公开发表。作者希望本书在促进语音技术的研究和开发方面起到抛砖引玉的作用。

本书适合于从事人工智能、模式识别、信息与控制、计算机应用的科技人员使用。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。

作者已竭尽全力来写好本书，但缺点和错误难以避免。恳请广大读者给予批评指正。

朱民雄

1991年12月于北京航空航天大学

目　录

试读结束，需要全本PDF请购买 www.ertongbook.com

第一章 概 述

自从本世纪70年代第一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，是一个多学科综合性研究领域，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一个重要部分，包括四种技术，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，本书仅涉及传统的四种语音技术。

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。本世纪 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一成果。1949 年贝尔实验室的研究人员研制成功第一个电合成器。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一个识别10个英文数字发音的实验系统。1960年Denes等人研究成功了第一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：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；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；用于输入语音与参考样本之间时间匹配的动态规划方法；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 6k 字节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公司推出另外一个产品ML-I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个单词和短语的音节词典。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128k位的ROM内存约330个单词和短语（语音持续约3～4分钟），数据传输率为1200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：

1. 与PC机兼容。
2. 开放式词汇表，用户可以方便地调整和扩充新词。
3. 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。
4. 具有很强的人机交互能力或友好用户接口。

另外一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：

1. 集成前人的研究成果，实现系统各个环节的优化。

2. 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。

3. 语音信号中提取多种特征，提高系统的识别率。

4. 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 0 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。它有三种工作方式：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。它有三个不同的词库模块，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5000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4k～8kHz，最大可达32kHz。存储时间有 8 秒、16秒、40秒、128秒等。存储器用1～4个256k位的DRAM。5～6V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 M位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

这一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。这一产品功能齐全，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75dB（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。

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

人类发出的语音波形是一种声压波。它是由图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五类：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。声带振动产生声音，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。

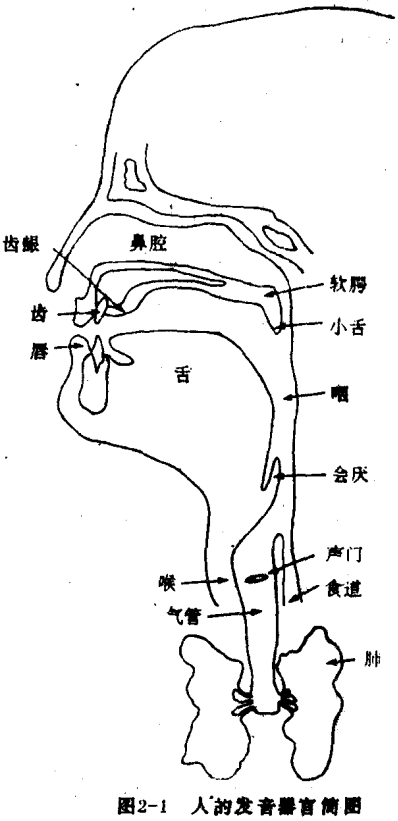


图2-1　人的发音器官简图

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50～250Hz，女性的范围约近于500Hz。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。

关于喉的发声机理有两种学说。一为张力学说，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。一为阵挛学说，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。

试读结束，需要全本PDF请购买 www.ertongbook.com

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。

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。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。声强大小取决于空气经过声带时的能量。供给的能量愈大，声带移动也愈大，产生的声波幅值也愈大。

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。声道是一谐振腔，它放大某一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。这就使讲话成为一连续变化的声波。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17cm，音速为340m/s，则在500、1500、2500Hz 产生谐振。这些共振峰常称为F1(约 500Hz)、F2(约 1780Hz)和 F3(约 2500 Hz)。这些共振峰的计算如下：

设男性成人的声道长 $L=17\text{cm}$，音速 $c=340\text{m/s}$，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幅射出去。λ_1、λ_2、λ_3 分别为第一共振峰频率的波长、第二共振峰频率的波长和第三共振峰频率的波长。

第一共振峰：

波长 $\lambda_1=4L=4\times0.17=0.68\text{m}$

频率 $F_1=\frac{c}{\lambda_1}=\frac{340}{0.68}=500\text{ Hz}$

第二共振峰：

波长 $\lambda_2=\frac{4}{3}L=0.2267\text{m}$

频率 $F_2=\frac{c}{\lambda_2}=1500\text{ Hz}$

第三共振峰：

波长 $\lambda_3=\frac{5}{3}L=0.136\text{m}$

频率 $F_3=\frac{c}{\lambda_3}=2500\text{ Hz}$

图 2-2 为 3 个共振峰在声道中传播的示意图。

虽然大部分元音的共振峰多于 3 个，但前 3 个共振峰已足够去表征和区别它们，其余的

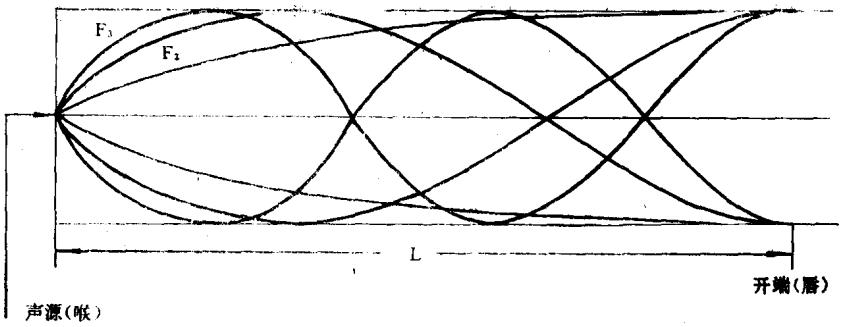


图2-2 共振峰在声道中传播

共振峰主要与个别讲话者的音质有关。声道的大小随不同讲话者而不同，因此共振峰频率与讲话者有密切关系。声道大小不仅影响共振峰频率，也影响其基频、所产生的声音和声道的整个外形。声道是辅音产生源，也是元音和辅音的谐振器。因而，辅音声如|l|、|m|，和|n|具有清晰的共振峰结构。图2-2假定声道为一均匀截面的声管，但实际声道是一根从声门延伸到唇的非均匀截面的声管，它的外形变化是时间的函数。

鼻腔——鼻腔在嘴的后部与口腔相通，有一称为软腭的活瓣守护它。靠软腭的帮助，可使空气经过鼻子排出人体外。鼻腔引起的共振峰谐振与声道相似。然而这些共振峰是固定的，因为没有办法去有效地改变鼻腔的大小。鼻腔与口腔协调工作产生声音。空气通过鼻腔而产生的语音叫作鼻音，如。|n|，|m|，|ŋ|。非鼻音由软腭完全堵塞通向鼻腔的通道，如|r|是非鼻音。当人患感冒时，由于空气通路受限或受阻，使产生的鼻音明显不同。

发声器官——不同声音的产生是由发声器官唇、齿、齿龈、舌、颌和面颊的组合而生成。唇的各种形状有利于产生元音和各种辅音，如|p|和|b|。齿与唇相结合产生语音。如发|f|音、首先使声带张开，把上齿放在下唇上，发声器官产生一次收缩，迫使气流通过声带。齿龈形成与舌尖的接触点，用于产生语音如|t|，|d|，和|n|。舌是非常灵活的肌肉器官。舌在口内不同部位的接触可以产生许多不同的语音。颌的运动改变口腔的大小和形状。面颊促使在口腔内形成压力，这种压力是正确产生某些语音如|p|和|b|所必需的。

综上所述，声音具有三个主要因素即音调、音强和音色。音调的高低和声带振动的频率有关，频率快则音调高，频率慢则音调低。声带振动的频率又决定于声带的长度、张力、厚薄和呼出气柱的强弱。一个训练有素的歌唱家，能精确地运用这些变化而发出准确的音调。音强的大小决定于振幅的大小和呼出气压的强弱。音色是由混入基音的泛音所决定，每个基音又都有其固有的频率和不同音强的泛音，使形成的每个声音各有其特殊的音色。所谓泛音乃是许多频率与基音频率成简单倍数的声音，如基音频率为100Hz，则泛音频率为200，300 400Hz等。每个人因其性别、年龄、喉部和声道构造的不同，产生泛音的成分也不同，故具有各不相同的音色，因此我们能够按口音分辨出每个说话的人。人类声音的音域随年龄增长而增加，成人约为两个8度音阶，对于具有高度训练的歌唱家可达2.5～3个8度音阶。一般谈话的声音常限于5度音阶之内，而不超过一个8度音阶。声音就其固有的音域和音乐特性可分为男低音（音域为81～325Hz），男中音（96～426Hz），男高音（122～580Hz）；女低音（145～690Hz），女中音（217～1024Hz）和女高音（256～1300Hz）等类型。

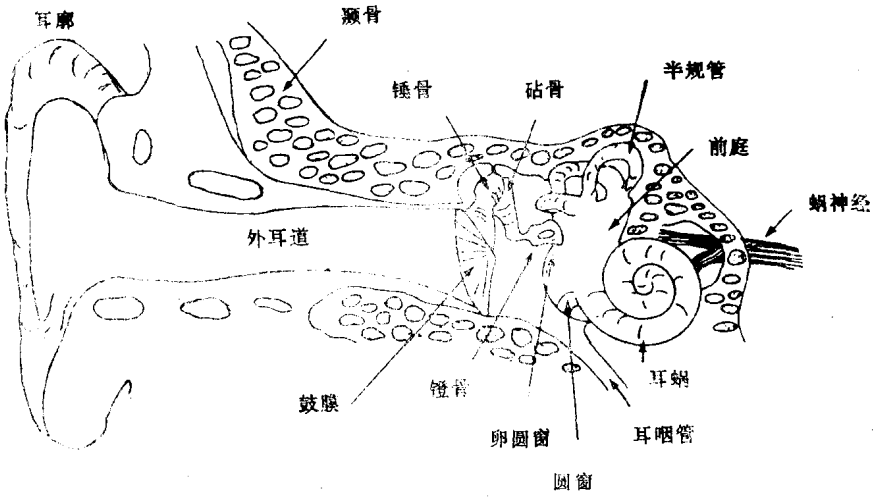


图2-3 人耳的解剖简图

2.1.2 语音接收过程生理学基础知识

人类接收语音由人耳来完成。对于人类的语音过程，语音分析表明，语音过程时会产生丰富的声学信息。然而，人类是不能感知所有这些信息的。例如，由同一个人以非常清晰的方法说同一个词，而语音分析时会产生不同的结果。但是，当我们听到这一声音时，会感知它们是同一个词。更为重要的是我们能感知不同讲话者所说的同一个词，尽管这个词具有明显不同的音调、语调、重音等。如果能感知所有这些信息，我们也不知道识别声音所需要的这些信息的量级。我们能感知语音需要多少信息仍是一个不能回答的问题。例如，我们知道共振峰F_1和F_2可用于区别不同的元音，而后面二个共振峰明显地告诉我们一个语音的音色。因而，我们还不知道如何识别不同讲话者声学上不同的词，“标准化”为同一个词。有一种确定如何识别不同的语音的方法是滤除某些频率，弄明白人耳是否能感知这一结果。为了弄清人类如何识别语音，简略介绍一些耳的生理学基础知识。

耳的主生理机能是听觉。人耳分为外耳、中耳和内耳三部分。外耳与中耳可从传导不良的介质即空气中收集声能，并将其传到内耳的淋巴（液体介质）中去，再激动耳蜗内的终器。此终器称螺旋器，又称柯替氏器，是听觉感受器。连到螺旋器上的蜗神经及前庭神经为感觉神经，即第八脑神经，又名位听神经。此神经经内耳门入颅，与脑干相连。语音信号通过外围的听觉系统后传到大脑。输入信号和第八脑神经接收到的信号是不同的。除此之外，这个过程我们知道得很少。

下面简要地讨论耳的生理特点及其对输入声波滤波和失真的方法。

外耳——从产生听觉的角度来看，外耳和中耳具有导音作用，故合称为导音系；内耳有感音作用，故称感音器。严格来说，内耳的感音作用实起自蜗神经的终器，即螺旋器，故内耳的淋巴系统也属导音系。外耳包括耳廓和外耳道。耳道直至鼓膜。耳道中充满空气，它本身的作用如一谐振器，使其谐振频率附近的频率有某些放大作用。这就导致原来产生的声音有某些失真。耳廓呈漏斗形，在集声过程中声音不致被歪曲。外耳道的共振作用可作如下

推算。外耳道平均长度以2.5cm计算。根据一端封闭的管腔，对波长为其4倍的声波能起最佳的共振作用的物理学原理，则2.5×4=10.0cm。3000Hz到4000Hz声波的波长为11.4～8.5cm。因此，外耳道对3000～4000Hz之间的频率有扩音作用。有的学者认为外耳道的共振作用能使此种频率的声波在鼓膜上声强级提高10dB。

中耳——中耳包括鼓室、咽鼓管等部分。借咽鼓管与鼻咽相通。鼓室为具有上、下、内、外、前、后六壁似一六面立方形小盒的含气空腔。鼓膜是鼓室外壁的主要组成部分。声波穿过耳道到鼓膜，声波的振动引起鼓膜的振动。然而，这仅限于鼓膜能振动的那些频率。但并不是声波中所有频率都会引起鼓膜振动，振动的上限频率约为20 000Hz。老年人稍低一些。鼓膜对所接收的一些频率滤波。鼓膜振动的响应频率范围并不是所有的人都一样。鼓膜的面积约为85～90mm^2，其有效振动面积55mm^2。鼓室中有三个听觉小骨（听小骨），即“砧骨”、“锤骨”和“镫骨”。锤骨系附于鼓膜，砧骨系附于锤骨内末端，另一端与镫骨相连，镫骨底部与内耳紧连。这些小骨的目的是检验已由鼓膜接受的声波，放大并传送这些声波到内耳。镫骨底面积为3.2mm^2，比鼓膜有效面积大17倍。鼓膜和听骨链连成一个整体，将振幅大、声压小的空气波转化为振幅小、声压大的内耳淋巴液波。因鼓膜表面的声压传到镫骨底时，此声压被提高17倍（因面积减小17倍）。中耳对所有频率的传送是不同的。在2 000Hz附近的频率它的传送工作最好，在高频和低频时其传送效率将降低。换言之，它也是一个滤波器，放大或衰减某些频率，使输入声波进一步失真。

内耳——内耳又称迷路，居于颞骨岩部之内。内耳结构比较复杂，包括耳蜗、半规管和前庭三部分。由镫骨底部直接连至一称为耳蜗的充满液体的空间。耳蜗位于前庭的前内方，形似蜗牛壳，为一长螺旋形骨管，共盘绕2.5～2.75周，非旋绕长度约为35mm。内耳充满着液体，并沿长度方向分成二路。内耳有二个开口通向中耳。镫骨与一个称为“卵圆窗”的开端相连。另一开端称为“圆窗”，盖着一层可使耳蜗的液体压力改变的弹性膜。当镫骨将振动传送到液体内，液体振动。隔膜及分路也吸收一些振动。由于这种结构，它沿其长度方向的不同点上，响应不同的谐振频率。被耳的各个不同器官感知的声音在耳器官结构内引起不同的物理变化。声音经耳器官的传送，到达内耳的终器（螺旋器）。终器有蜗神经与脑干相连。传到大脑的声音是被稍微放大或衰减，且滤掉某些频率。这一声音信息传送到脑部作进一步处理，而我们对这一处理了解甚少。耳蜗的感音生理有多种学说。远在100年以前，听觉学说就想研究解决这样的问题，即耳蜗是一个声音分析器，还是一个换能器，而把分析任务交给中枢神经去完成？然而问题至今没有解决。

最后再说明一下声音传入内耳的途径。声音可由两条途径传入内耳。一为空气传导（气导），空气振动的声波由耳廓收集，经外耳道而抵达鼓膜，使鼓膜随着振动。振动的结果使听骨链和鼓室内的空气也发生振动。听骨链的振动经卵圆窗激动前庭淋巴，变为液波，液波振动基底膜，使位于基底膜上的螺旋器受到刺激，将冲动经听神经传至中枢而产生听觉。另一为骨传导（骨导），声波直接经过颅骨传导，使外淋巴发生相应振动，再激动耳蜗的终器产生听觉。骨导虽发生于气导的同时，但经骨导进入耳蜗的声能殊微，实无重要意义。因此，声音传入内耳的途径以气导为主。

§2.2 语音学基础知识

世界上有许多种语言，其中有些语言的文字表示与发音是不同的。因此学习者必需掌握语言的表音法，表音法是指用文字或印刷符号标出某一语言的音。汉语和英语都是这样的语言。掌握语言的语音学知识有助于学会语言的表音法。语言的语音学知识是计算机语音分析的基础，而语音分析又是计算机语音合成和识别的基础。因此，每一个从事计算机语音技术的科技人员均要学好语音学基础知识。这里以英语为例来阐述语音学基础知识。

2.2.1 词的分段特点

英语中将语音分为元音和辅音两大类。元音的特点是声道没有闭止或阻塞。所有的元音都是有声音的。元音是一类音素，它在词和其他语言结构中具有类似的位置。一般说，元音有五个（a、e、i、o、u）。而从语音学角度来说，就有19个元音/双元音音素。它们和24个辅音一共43个音素，足以正确地描述英语。元音声的产生相对于辅音声有根本的不同。最重要的不同点在于：产生辅音声时，从肺部出来的气流由于发音器官的接触而受到某些限制，或者是由于声道变窄而受到限制，因而所有的辅音并不都是有声的。这就使元音和辅音有不同的分类系统。已经证明：发音器官的位置大大地有助于语音的声学分析。在发音器官的位置和作用的基础上，语音学家已提出许多不同音素分类方法。

不同的元音声可根据舌的位置、肌肉紧张程度、舌尖的卷曲和圆唇的程度来分类。

舌的位置——在英语中，由舌在口腔内的3个位置来产生元音。此时伴随着舌在水平方向的运动和涉及舌的上下位置。口腔的外形影响产生元音的质和量。产生元音舌的位置有在口的前、中和后3种，在不同元音时改变谐振腔。这是舌在水平方向的3个位置，它同时也沿垂直方向运动。产生8个英语元音，舌有9种位置。表2-1为仅由舌的位置的元音分类。

另有6个元音与上述8个元音不同。它们需要附加特点才能明显地分类。表2-2给出了这些元音，而在下面会讲到其分类。

表2-1 仅由舌的位置的元音分类

音素	例 词
|i|	eat, beet, bee
|u|	ooze, boot, shoe
|e|	age, vacate, say
|æ|	add, bat
|ə|	about, potato
|ʌ|	up, but
|o|	open, rotate
|a|	otter, hot, spa

表2-2 仅由舌的位置不能分类的元音

音素	例 词
|I|	it, bit, hippie
|ɛ|	end, bet
|ɔ|	urdane, purport, sister
|ʊ|	put
|ɔ|	awful, bought, paw
|ɜ^|	irk, bird, burr

肌肉紧张程度——由紧张程度的附加发音特点可区别元音，如|i|和|I|，|u|和|v|。“紧”元音由附加肌肉紧张而产生，而“松”元音没有肌肉紧张而产生。现考虑在词“wood” |wud|和词“would” |wʊd|中的两元音|u|和|v|。元音|u|发音时有紧张特点，而元音|v|

是松元音，无此特点。紧元音是|i|、|e|、|ə|、|ɔ|、|u|和|o|，它们对应的松元音是|I|、|ɛ|、|(ɛ)|、|(ɜ^)|、|ʊ|和|ɔ|。元音|æ|、|ʌ|的紧张特点是中性的。所有14个元音都有清楚的肌肉紧张特点。

舌尖卷曲（卷舌）——除了"u·bane"中的|ɔ|和"bird"中的|ɜ^|而外，所有元音均是非卷舌的。前二元音的发音舌尖需要卷曲，故为卷舌音。

圆唇程度——14个英语元音中|u|、|ʊ|、|o|、|ɔ|和|a|为圆唇元音，发音时要附加圆唇。非圆唇元音|i|、|I|、|e|、|ɛ|和|æ|不需圆唇而发音。中性元音是|ə|、|ɔ|、|ɜ^|和|ʌ|。圆唇元音发音时，舌的位置在后面。非圆唇元音发音时，舌的位置在前面。当元音后跟以鼻辅音时，英语元音也存在鼻音。

一般来说，英语有21个辅音。而从语音学角度来说，英语有24个辅音。辅音有两种常用的分类方法。按发音位置辅音可分为6个子群。按发音方式辅音有5个子群。

按发音位置分类——双唇音音素全由唇产生，且包括了上下两唇，因而得此名称。|p|，|b|，|m|，和|w|是双唇辅音。唇齿辅音的产生是下唇和上齿相接触，如辅音|f|和|v|。唇齿辅音和双唇辅音合为一类，称唇辅音。舌齿辅音的产生是舌在上下齿之间，如"this"中的|ð|和"thank"中的|θ|。有时它们称作齿间辅音。齿龈音的产生是舌和齿龈相接触。齿龈是舌的最自然接触点，也许在世界上许多语言中用这种方法产生高频声音。例如，产生英语声音|t|、|d|和|n|；而声音|s|、|z|、|l|和|r|是舌紧贴齿龈而产生的。上述4子群中，谐振腔的大小在阻塞点后的要大于阻塞点前的。因此，这4子群有时称作"前辅音"。腭音的产生是舌体与腭相接触。下面是七个腭音及其相应词的例子："shut"和"sugar"中的|sh|，"church"、"trench"中的|ch|，"bridge"中的"dg"等。应该指出，发腭音时限制点后的谐振腔减小了，而阻塞点前的腔增大了。这些辅音称作"后辅音"。软腭音的产生是舌接触或接近软腭。如语音|k|、|g|和|h|及"sing"中的"ng"等。上述例子仅对英语而言，对其他语言要作某些改变。据说在世界上所有的语言中，人能够做到形成26个不同的发音接触点。这意味着存在着比我们上面讨论过的更多的音素和双元音。

辅音可以根据其发音方法而分类。这种分类是很清楚的，它们可以唯一地区别出一定的音素类。下面是根据发音方式的辅音分类。

共振音/阻塞音——根据发音所需的声道阻塞程度来区分所有的辅音。共振音仅需要最小的阻塞而阻塞音要求相当的阻塞量。

鼻音/口腔音——根据所用谐振腔来区分语音。口腔语音仅用口腔（不用鼻腔），鼻音用包括鼻腔的全声道。

闭止音/摩擦音——这二种辅音均要求声道阻塞，因而分类为阻塞音。然而，闭止音要求声道在一定点上完全关闭，而摩擦音不必这样。闭止音也叫"爆破音"，这一术语的意思是在一闭止后气流"爆破放出"。通常闭止发生在词或音节的起始时，因而此术语"爆破"仅对初始闭止或音节是合适的。

丝音/非丝音——丝音由突出的嘶嘶声来区分，如|s|。

浊音/清音——浊音由出现声道振动来区分。所有的共振音和所有的元音是浊音。在这里的分类中，这种浊音是不用的。然而，并不是所有的阻塞音都是浊音，因而，这一分类用于细分阻塞音。

试读结束，需要全本PDF请购买 www.ertongbook.com