高等院校
通信与信息专业规划教材

现代语音处理技术及应用

张雄伟　陈　亮　杨吉斌　编著

机械工业出版社
CHINA MACHINE PRESS

高等院校通信与信息专业规划教材

现代语音处理技术及应用

张雄伟　陈　亮　杨吉斌　编著

机 械 工 业 出 版 社

本书从人类的发声机理和听觉机理出发，全面系统地介绍了现代语音信号处理的基础、原理、方法与应用。首先介绍了语音信号的基本性质和数学模型；详细阐述了短时时域处理技术、变换域分析、线性预测分析、矢量量化的基本原理与方法；重点介绍了语音编码、语音识别、语音合成和语音增强等语音处理的几项最重要的技术；最后介绍了语音通信应用中的几个关键技术和实时语音处理系统设计的基本方法。着眼于语音信号处理的新发展，本书还对信号处理领域的小波、混沌、分形以及人工神经网络等新技术新方法在语音信号处理中的应用进行了讨论。附录部分给出了语音处理有关技术的理论推导及一些实用的 C 程序和 MATLAB 程序的实例，供相关人员学习应用时参考。

本书内容广泛，重点突出，原理阐述深入浅出，注重理论与实际应用的结合，可读性强。本书可作为高等院校通信工程、电子工程、信息工程等专业高年级本科生和信号与信息处理、通信与信息系统等学科研究生的教材，也可供语音处理和信息技术研究的科研及工程人员参考。

图书在版编目（CIP）数据

现代语音处理技术及应用/张雄伟等编著. —北京：机械工业出版社，2003.8

高等院校通信与信息专业规划教材

ISBN 7-111-12795-1

Ⅰ.现… Ⅱ.张… Ⅲ.语音处理—高等学校—教材 Ⅳ.TN912.3

中国版本图书馆 CIP 数据核字（2003）第 067005 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策　　划：胡毓坚

责任编辑：孙　业

责任印制：路　琳

北京蓝海印刷有限公司印刷·新华书店北京发行所发行

2003 年 8 月第 1 版·第 1 次印刷

787mm×1092mm 1/16·20.75 印张·512 千字

0 001—5 000 册

定价：29.00 元

高等院校通信与信息专业规划教材
编委会名单

（按姓氏笔画排序）

出 版 说 明

为了培养21世纪国家和社会急需的通信与信息领域的高级科技人才，为了配合高等院校通信与信息专业的教学改革和教材建设，机械工业出版社会同全国在通信与信息领域具有雄厚师资和技术力量的高等院校，组成阵容强大的编委会，组织长期从事教学的骨干教师编写了这套面向普通高等院校的通信与信息专业规划教材，并且将陆续出版。

这套教材将力求做到：专业基础课教材概念清晰、理论准确、深度合理，并注意与专业课教学的衔接；专业课教材覆盖面广、深度适中，不仅体现相关领域的最新进展，而且注重理论联系实际。

这套教材的选题是开放式的。随着现代通信与信息技术日新月异地发展，我们将不断更新和补充选题，使这套教材及时反映通信与信息领域的新发展和新技术。我们也欢迎在教学第一线有丰富教学经验的教师及通信与信息领域的科技人员积极参与这项工作。

由于通信与信息技术发展迅速而且涉及领域非常宽，这套教材的选题和编审难免有缺点和不足之处，诚恳希望各位老师和同学提出宝贵意见，以利于今后不断改进。

机械工业出版社
高等院校通信与信息专业规划教材编委会

前　言

语音是人类相互交流和通信最方便快捷的手段。如何高效地实现语音传输、存储或通过语音实现人机交互，是语音信号处理领域中的重要研究课题。语音信号处理涉及数字信号处理、语言学、语音学、生理学、心理学、计算机科学以及模式识别、人工智能等诸多学科领域，是目前信息科学技术学科中发展最为迅速的一个领域。

近20多年来，语音处理技术取得了一系列重大进展，语音编码、语音合成、语音识别和说话人识别等方向的研究成果不断推出；同时，微电子技术的迅猛发展和数字信号处理(DSP)芯片性能的不断提高，为实时实现更高复杂度的高性能语音处理算法提供了可能。目前市场上已有不少语音处理的应用产品，并且不断有许多新产品推出，语音处理技术的应用前景和市场潜力十分巨大。

本书是作者以近几年来为本科生和研究生讲授“语音信号处理”课程的讲义为基础，结合主持开发过的多项语音处理科研项目，并参考相关文献资料和最新动态编写而成。全书系统介绍了语音信号处理的基本理论、方法和相关的应用领域，并介绍了目前的研究现状和学科发展趋势，内容通俗易懂、深入浅出。

全书共分12章。第1章是绪论；第2章介绍语音处理的基础知识；第3章着重讨论语音信号的时域分析技术；第4章讨论语音信号的变换域分析技术；第5章介绍现代语音处理中最重要的一项技术——语音信号线性预测分析技术；第6章重点阐述语音的矢量量化技术；第7章介绍语音编码技术，包括波形编码、参数编码和混合编码，并介绍常用的语音编码标准及语音编码的质量评估方法；第8章讨论语音识别、说话人识别的基本原理以及典型的语音识别系统；第9章介绍了语音合成技术；第10章是语音增强技术；第11章介绍语音通信应用中的几个关键技术；第12章介绍了基于DSP芯片技术的实时语音处理系统。为了加深理解，拓宽思路，在某些章中结合实际给出了一些具体的应用实例。

本书由张雄伟、陈亮、杨吉斌编写。张雄伟编写了第1、2、11、12章，陈亮编写了第3、4、5、7章，杨吉斌编写了第6、8、9、10章，邹霞和贾冲给予了很大帮助，曹铁勇、黄忠虎、王金明提出了具体建议。全书由张雄伟、陈亮审校和统稿。在本书的策划和编写过程中，南京邮电学院的杨震教授和东南大学吴镇扬教授提出了宝贵的建议，解放军理工大学通信工程学院教保科的许晔峰和乐超为本书的出版做了许多工作，在此一并向他们表示衷心的感谢。

本书得到解放军理工大学通信工程学院的资助。

由于作者水平所限，疏漏和错误之处在所难免，请广大读者批评指正。

编　者

目　录

第1章 绪 论

1.1 概述

语音是人类相互之间进行交流时使用最多、最自然、最基本也是最重要的信息载体。在高度信息化的今天,语音处理的一系列技术及其应用已经成为信息社会不可或缺的重要组成部分。

语音的产生是一个复杂的过程,包括心理和生理等方面的一系列动作。当人需要通过语音表达某种信息时,首先是这种信息以某种抽象的形式表现在说话人的大脑里,然后转换为一组神经信号,这些神经信号作用于发声器官,从而产生携带信息的语音信号。

语音信号处理的研究,起源于对发声器官的模拟。1939 年,美国人 H.Dudley 展出了一个简单的发声过程模拟系统,以后发展成为声道的数字模型。利用该模型可以对语音信号进行各种频谱及参数的分析,同时也可根据分析获得的频谱特征或参数变化规律,合成语音信号,实现机器的语音合成。

目前,对语音信号进行研究一般都基于语音信号的数字表示,因此,语音信号的数字表示是进行语音信号数字处理的基础。语音信号数字化的理论依据是我们熟知的采样定理,即只要采样频率足够高,就可以用时域上周期抽取的样点来表示一个带限信号。语音信号的离散表示基本上可以分为两大类:波形表示和参数表示。波形表示仅仅是通过采样和量化的过程保存模拟语音信号的"波形",而参数表示则是把语音信号表示成某种语音产生模型的输出。为了得到参数表示,首先必须对语音进行采样和量化,然后再进一步处理以得到语音产生模型的参数。语音模型的参数一般可分为两大类:一类是激励参数;另一类是声道参数,如图 1-1 所示。

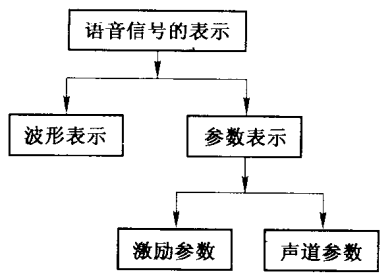

图 1-1 语音信号的表示方法

由于语音的特殊作用,人们历来十分重视对语音信号和语音通信的研究。社会的进步对语音通信提出了更高的要求,需要更高的语音质量和更低的数码率,从而推动了语音编码技术的发展。而自动控制和计算机科学的发展又要求用语音实现人与机器的信息交流,要求机器能听懂人说话和模仿人说话,甚至还要能辨别说话人是谁,这又推动了语音识别和语音合成技术的研究,使语音处理技术得到迅速的发展。语音编码、语音识别、说话人识别、语音合成等技术的基础都是对语音信号特征的认识,都要利用数字信号处理的一些基本技术来分析和处理语音信号,而更深层次的发展涉及人的发音和听觉机理,与生理学、语言学甚至心理学有关。

语音信号数字处理是一门涉及诸多学科的交叉学科,它以生理学、心理学、语言学以及声学等学科为基础,以信息论、控制论、系统论的理论作指导,通过应用信号处理、统计分析、模式

识别等现代技术手段而发展形成的一门综合性学科。20 世纪 80 年代以前，线性预测编码(LPC)技术是语音信号处理研究领域最重要的研究成果。80 年代以后，分析合成技术、矢量量化技术、隐马尔可夫模型(HMM)等极大地推动了语音编码、语音识别技术发展。90 年代以后，神经网络、小波分析、分形及混沌等新技术在语音处理领域的应用将语音信号处理的研究提高到了一个新的水平。

尽管语音处理的研究已经经历了几十年的历史，并已取得许多成果，但语音处理的研究仍然蕴涵着巨大的潜力，还面临着许多理论和方法上的实际问题。例如，在语音编码技术方面，能否在极低速率或甚低速率下取得满意的语音质量；在语音识别方面，连续语音的分割、大词汇量语音识别及识别任何人的语音等方面目前尚没有十分理想的办法。在语音理解方面，关于语义信息的定性描述和定量估计等，都还没有统一的计算方法。所有这些都是语音处理领域今后研究的重要方向。

对语音信号进行研究可以有不同的方法，本书着重从数字信号处理的角度来研究语音信号的各种处理方法。

1.2 语音处理的研究方法

语音处理主要从基础理论、算法实现及实际应用等几个方面来研究。对语音处理的基础理论及各种处理算法的研究主要包括以下两个方面。

一是从语音产生和语音感知来研究。语音产生的研究涉及大脑中枢的言语活动如何转换成人发声器官的运动，从而形成声波的传播。语音感知的研究涉及人耳对声波的收集并经过初步处理后转换成神经元的活动，然后逐级传递到大脑皮层的语言中枢。语音产生和语音感知方面的研究与语音学、语言学、心理学和神经生理学等学科紧密相关。目前，对于整个语言链的物理层(包括发声器官和人耳的功能)已经研究得比较透彻，而对于神经元活动和大脑语言中枢的工作原理还有待今后进一步研究。

二是将语音作为一种信号进行处理。20 世纪 60 年代形成的一系列数字信号处理方法和算法，如数字滤波器、FFT 等与语音信号处理紧密联系。后来出现的线性预测编码技术成为语音信号最有效的处理方法之一，广泛应用于语音分析合成及各个语音应用领域。80 年代出现的分析合成法、码激励线性预测(CELP)、矢量量化(VQ)以及隐马尔可夫模型(HMM)等一系列算法和模型极大地推动了语音编码和语音识别技术的研究。

本书对第一方面内容仅做简要介绍，重点介绍的是第二方面的内容。

1.3 语音处理的应用

语音处理的应用非常广泛，最基本的应用就是语音的数字传输，即将语音进行数字化后在数字通信系统中进行传输，以实现数字语音通信。图 1-2 列出了语音数字处理的一些典型应用。

下面简要介绍一下这些应用。

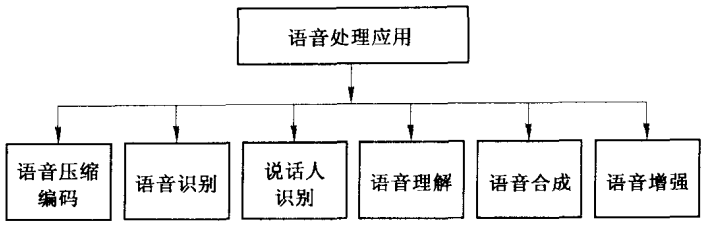

图 1-2　语音数字处理的典型应用

1.3.1　语音压缩编码

语音压缩编码是语音数字处理最重要的一种应用。语音压缩编码的目的是用尽可能低的比特率来获得尽可能高的合成语音质量。实现语音压缩编码(特别是中低速率)的设备通常称为声码器。虽然光纤通信和微波通信等系统可以提供很宽的频带,但在很多情况下仍然需要压缩语音编码速率以节省频带。一方面压缩编码后可以在有限带宽的信道上传输多路语音,提高信道的利用率;另一方面可以在窄带的模拟信道(如短波)上传输数字语音。特别是在军事通信系统等需要复杂加密的应用场合,声码器具有不可替代的作用。此外,语音的数字存储、语音应答等也是语音压缩编码的重要应用。在语音压缩编码技术中,线性预测、矢量量化、码本激励等是最重要的几种实现技术。根据语音压缩编码的采样率,可以分为窄带(电话带宽300～3400 Hz)语音压缩编码、宽带(7 kHz)语音压缩编码和 20 kHz 的音乐带宽压缩编码。窄带语音压缩编码的采样率通常为 8 kHz,一般应用于语音通信中。宽带(7 kHz)语音压缩编码的采样率通常为 16 kHz,一般用于要求更高音质的应用中,如会议电视。而 20 kHz 宽带主要是适用于音乐数字化,采样频率高达 44.1 kHz。

近几十年来,语音编码技术发展非常迅速。以窄带语音编码为例,自 20 世纪 70 年代推出 64 Kb/s PCM 语音编码国际标准以来,已相继有 32 Kb/s ADPCM、16 Kb/s LD-CELP、8Kb/s CS-ACELP 等国际标准推出。而地区性或业务性的标准也有不少,如第二代移动通信系统中的语音编码,美国国防部制定的 4.8 Kb/s 和 2.4 Kb/s 保密电话标准等。

目前,在 2.4Kb/s 以上的编码速率,合成语音质量已得到人们的认可,并已广泛应用。未来的研究重点是突破 2.4 Kb/s 以下极低速率的语音编码技术和算法。

1.3.2　语音识别

语音识别的作用是将语音转换成等价的书面信息,也就是让计算机听懂人说话。目前语音识别已经成为语音数字处理研究领域中的重点和难点技术。语音识别可以有许多分类方法,例如,根据语音识别对象来划分,可以分为孤立词识别、连续语音识别等;根据词汇量来划分,可以分为小词汇表(100 个词以下)、中词汇表(100～500 个词)、大词汇表(500 个词以上)语音识别等;根据对说话人的要求来划分,可以分为特定说话人(Speaker Dependent)语音识别、多说话人语音识别和非特定说话人(Speaker Independent)语音识别等。

语音识别虽然从原理上看实现并不困难,但在实际实现时遇到的困难很多。例如,发音的多变性、不同人发同一个音、同一个人在不同的条件下发同一个音等都会有不同的发音参数;发音的模糊性,在实际的连续语音流中语音声学变量与音素变量之间不存在一一对应关系;语音流中变化多端的音变现象,这些音变对人类的听觉系统来说很容易辨认,但机器识别却很不

容易。

语音识别的应用很广，声控打字、用声音控制计算机等，如将语音识别与语音合成结合起来就可以实现甚低比特率的语音通信系统。

目前，一些中、小词汇量的孤立词或连续语音识别系统已进入市场。目前研究的重点是实现大词汇表、非特定人的连续语音识别系统，它可用于人机直接对话、语音打字机以及两种语音之间的直接通信等一系列重要场合。

1.3.3 说话人识别

说话人识别的作用是根据语音辨别说话人，广义的语音识别也包括说话人识别。但说话人识别并不注意语音信号中的语义内容，而是希望从语音信号中提取出人的特征，即根据语音判别说话人是谁。语音信号既载有说话人的语言信息，同时也载有说话人本身的特征信息。每个人的发音器官都有自己的特征，说话时也都有自己的特殊语言习惯。在分析语音信号时，可以提取说话人的个人特征，从而有可能识别说话人是谁。在语音识别时，要消除说话人的个人特征，以免影响识别的准确率；而在研究说话人识别时，则要专门研究人的特征，从语音信号中分析和提取个人特征，去除不含个人特征的语音信息。

说话人识别包含两个方面：1)说话人确认；2)说话人辨认。前者是确认说话人的身份，说话人说一句或几句测试语句，经处理后获取的特征参数与储存的特定人语音的参数比较，作出“是与否”的判决。后者是要辨认待识别的来自若干人中的哪一位，要将待识语音与每一位的语音比较，找出距离最近的语音所对应的说话人。从语音信号处理的角度来看，二者基本上是相同的，都需要确定选用的参数和计算距离的准则。前者需确定“是与否”的门限，后者需与待识语音比较它们各自的距离。比较的方法与识别语音的方法相类似。参数的选择原则，一是要能反映说话人的个性，二是要兼顾识别率和复杂程度。比较简单的特征参数是基音和能量，也可以用 LPC 参数与共振峰，但计算量稍大。也有用语谱图来识别的，称为“声纹”。

然而由于语音是动态的，它和说话人所处的环境、情绪和身体状况关系很大。一个人在不同时间不同情况下说同一句话，差异也不一定比不同人小，不像“指纹”是静态的、绝对的。在现阶段还需结合识别人员的经验以提高识别的准确率，这方面的研究还在继续。还有一些识别难度更大，但更有实际价值的领域。如：1)用通过电话信道的语音进行“说话人识别”。由于电话频带窄、有失真、噪声大，不同信道条件各异，识别十分困难。但这方面的研究具有重要的实际价值。2)在“辨认”说话人时，语句往往不能规定，在没有指定语句条件下的识别也较困难。必须有更多的样本用作训练和测试，以降低误识率。这类无指定测试语句的识别称为“与文本无关”的识别。而在有指定语句条件下进行的识别称为“与文本有关”的说话人识别。

1.3.4 语音理解

语音理解是利用知识表达和组织等人工智能技术进行语句自动识别和语意理解。与语音识别的主要不同是对语法和语义知识的充分利用程度。由于人们对语音具有广泛的知识，可以对要说的话具有一定的预见性，所以人对语音具有感知分析的能力。依靠人对语言和谈论的内容多具有的广泛知识以及利用知识提高计算机理解语言的能力，是语音理解研究的核心。

利用理解能力，不仅可以排除噪声的影响，理解上下文的意思并能用它来纠正错误，澄清不确定的语义，而且能够处理不合语法或不完整的语句。一个语音理解系统除了包括原语音

识别所要求的部分之外，还必须增加知识处理部分。知识处理包括知识的自动收集、知识库的形成、知识的推理与检验等。当然还希望能自动地作知识修正的能力。因此，语音理解可以认为是信号处理与知识处理的产物。语音知识包括音位知识、音变知识、韵律知识、词法知识、句法知识、语义知识以及语用知识。这些知识涉及语音学、汉语语法、自然语言理解以及知识搜索等许多交叉学科。

实现完善的语音理解系统是非常困难的，然而面向特定任务的语音理解系统是可以实现的，例如飞机票预售系统、银行业务、旅馆业务的登记及询问系统等。

1.3.5 语音合成

语音合成的目的就是让计算机说话。

最简单的语音合成应当是语音响应系统，其实现技术非常简单。在计算机内建立一个语音库，将可能用到的单字、词组或一些句子的声音信号编码后存入计算机，当键入所要的字、词组或句子代码时，就能调出对应的数码信号，并转换成声音。

按规则的文字-语音合成系统是将文字转换成语言，让计算机模仿人来朗读文本。系统具有以下作用：有一存储基本语音单元的音库；当用各种方式输入文字信息时，计算机能将文字内容按照语言规则，转换成由基本音元组成的序列；按说话时音元连接的规则控制音元序列，输出连续自然的声音。这种系统也称“文-语转换系统”(TTS，Text To Speech)。

建立音库时语音单元的选择是一个很重要的问题。因为一种语言的音素通常只有几十个，采用音素作为音元可以降低存储容量，但用音素合成语音非常复杂，而且自然度较差。因此一般认为，汉语中采用音节作为音元比较合适，因为汉语中一个音节就是一个字的音，汉语中只有 412 个无调音节，形成音库比较适中。也可以用单字和词组作为音元，但一个字不能只存一种发音，因为汉语是多音字，字的发音与上下文有关，只有存储与上下文关联的几种发音，使用时按上下文关系调用，合成的语音次才能比较自然，这就要求有很大的存储容量。

系统中的“规则”有两层含义：一是文字变语言，如“。”要置换成“句号”；另一层是还要按照复杂的语音规则和上下文的关系决定音调、语气、重音、音长、停顿、过渡等，组成发音控制参数序列。

要使文-语转换系统合成出高质量的语音，不仅要掌握语音信号的数字处理技术，而且要有语言学知识的支持。

更高层次的合成是“按概念或意向到语音的合成”。要将“想法、意向”组成语言并变成声音，就如大脑形成说话内容并控制发声器官产生声音一样。

1.3.6 语音增强

在实际的应用环境中，语音都会不同程度地受到环境噪声的干扰。语音增强就是对带噪语音进行处理，降低噪声的影响，改善听觉的效果。有些语音编码和语音识别系统在无噪声或噪声很小的环境中性能很好，但当环境噪声增大时，性能却急剧下降。因此，最大程度地去除噪声，改善听觉效果，也是语音编码和语音识别等系统必须解决的问题。

实际语音遇到的干扰可能有以下几类：1)周期性噪声，如电气干扰，发动机旋转引起的干扰等，这类干扰在频域上表现为一些离散的窄峰；2)冲激噪声，如电火花、放电产生的噪声干扰，这类干扰在时域上表现为突然出现的窄脉冲；3)宽带噪声，这是指高斯噪声或白噪声一类

的噪声，其特点是频带宽，几乎覆盖整个语音频带；4)语音干扰，如话筒中同时进入多个人的声音，或者在传输时遇到串音引起的语音噪声。

对于上述各种不同类型的噪声，语音增强的方法也是不同的。例如，周期性噪声可以用滤波的方法滤除。冲激噪声可以通过相邻的样本值，采取内插方法将其去除，或者利用非线性滤波器滤除。宽带噪声是一种难以滤除的干扰，因为它与语音具有相同的频带，在消除噪声的同时将不可避免地影响语音的质量，现在常用的方法有谱减法、自相关相减法、最大似然估计法、自适应抵消法等。语音相互干扰也是很难消除的，一般可以采用自适应技术跟踪某个说话人的特征的方法来消除。

1.4 本书的内容与组织

本书是按照基础知识——技术与算法——实现的主线展开的。第一部分是语音处理的基础知识，包括：语音信号处理基础(第 2 章)；第二部分是本书的重点，介绍语音处理的技术和算法，包括：语音信号的时域分析(第 3 章)，语音信号的变换域分析(第 4 章)，语音信号线性预测分析(第 5 章)，矢量量化(第 6 章)，语音编码(第 7 章)，语音识别(第 8 章)，语音合成(第 9 章)和语音增强(第 10 章)；第三部分是语音处理的实用技术，包括：语音通信应用中的关键技术(第 11 章)和语音处理的实时实现(第 12 章)。

1.5 习题

1. 语音信号可以用什么方式表示？
2. 语音处理的典型应用有哪些？
3. 解释下列名词：语音压缩编码、语音识别、说话人识别、语音理解、语音合成、语音增强。