经典译丛

信息与通信技术

Theory and Applications of Digital Speech Processing

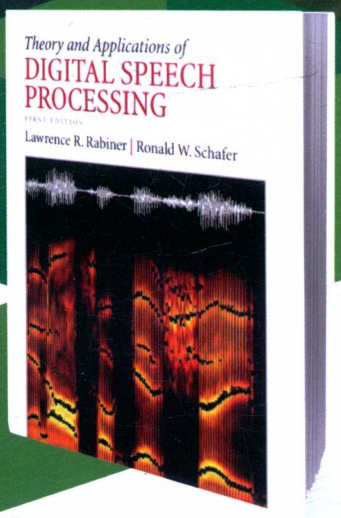
PEARSON

数字语音处理理论与应用

Theory and Applications of Digital Speech Processing

【美】 Lawrence R. Rabiner
Ronald W. Schafer 著

刘 加 张卫强 何 亮 路 程 等译



中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

经典译丛·信息与通信技术

数字语音处理
理论与应用

Theory and Applications of Digital Speech Processing

[美] Lawrence R. Rabiner
Ronald W. Schafer 著

刘 加 张卫强 何 亮 路 程 等译

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书是作者继1978年出版的经典教材《语音信号的数字处理》之后的又一著作，全书除有简练精辟的基础知识介绍外，系统讲解了近30年来语音信号处理的新理论、新方法和在应用上的新进展。全书共14章，分四部分：第一部分介绍语音信号处理基础知识，主要包括数字信号处理基础、语音产生机理、（人的）听觉和听感知机理，以及声道中的声传播原理；第二部分介绍语音信号的时、频域表示和分析；第三部分介绍语音参数估计方法；第四部分介绍语音信号处理的应用，主要包括语音编码、语音和音频信号的频域编辑、语音合成、语音识别及自然语言理解。

本书可供高等院校通信、电子、计算机等专业作为研究生和本科生的教材，也可供相关科研和工程技术人员参考，是一本既有系统的基础理论讲解，又有最新研究前沿介绍并密切结合应用发展的教材。

版权贸易合同登记号　图字：01-2010-5783

图书在版编目（CIP）数据

数字语音处理理论与应用/（美）拉比纳（Rabiner，L. R.），（美）谢弗（Schafer，R. W.）著；刘加等译.
北京：电子工业出版社，2016.1
书名原文：Theory and Applications of Digital Speech Processing
ISBN 978-7-121-27590-6

I. ①数…　II. ①拉…　②谢…　③刘…　III. ①语音数据处理－高等学校－教材　Ⅳ. ①TN912.3

中国版本图书馆 CIP 数据核字（2015）第 273596 号

策划编辑：马　岚
责任编辑：谭海平
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1 092　1/16　印张：42.5　字数：1196 千字　彩插：2
版　　次：2016 年 1 月第 1 版
印　　次：2016 年 1 月第 1 次印刷
定　　价：128.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

译 者 序

语音信号处理是一门古老而新颖的学科，说它“古老”是因为它与数字信号处理同时代产生，说它“新颖”是因为它一直经历着令人激动的变革和挑战。Lawrence R. Rabiner 教授作为这些变革的亲历者和大师级人物，有着深刻的切身体验，他的著作，如 1978 年他与 Ronald W. Schafer 教授合著的《语音信号数字处理》和 1993 年他与 Biing-Hwang Juang 教授合著的《语音识别基本原理》，也成为了语音信号处理领域的经典和必备读物。2010 年，在清华大学电子工程系朱雪龙教授的推荐下，电子工业出版社希望我们完成 Rabiner 教授和 Schafer 教授的新作《数字语音处理理论与应用》一书的翻译工作，我们欣然接受了翻译任务。然而，翻译的过程是艰辛的，为了能够对原文有比较准确的翻译表述，我们经历了无数个不眠之夜，历时五载，终于完成了初稿。在此期间，由于机器学习（尤其是深度学习）、听觉感知、听觉场景分析等理论和技术的发展，语音信号和信息处理技术经过一段平缓发展期后，又开始生机盎然，语音识别、说话人识别、语种识别、语音增强、语音和音频编解码、自然语言处理等技术都有新的创新，其系统性能也有显著提升。语音相关的产品也如雨后春笋般地涌现。在此时机下，我们期待此书的翻译出版能对国内语音界的科研人员，以及本科生和研究生的专业教学有所帮助。

本书原著结合自己的科研实践对数字语音信号处理的基本原理和应用进行了深入分析，既有理论深度，又通俗易读。内容分为四个层次逐级展开：第一个层次介绍语音信号处理基础知识，主要包括数字信号处理基础、语音产生机理、人的听觉和听感知与声道中的声传播；第二个层次介绍语音信号的时频表示，主要包括时域表示、频域表示、倒谱及同态处理和线性预测分析；第三个层次介绍语音参数估计算法，主要包括静音检测、清浊判断、基音和共振峰估计等；第四个层次介绍语音信号处理的应用，主要包括语音编码、语音和音频频域编码、语音合成、语音识别和自然语言理解。除了深入浅出的讲解外，书中还附有大量生动的插图，各章之后还附有精心设计的习题和 MATLAB 练习，以便读者对基础知识和基本方法深入理解和灵活应用。

本书能够得以完成，要特别感谢清华大学的朱雪龙教授，他不但为我们和出版社牵线搭桥，而且一直关心着我们的翻译工作；另外他于 1983 年牵头翻译的《语音信号数字处理》也为本书提供了诸多宝贵的参考和基础。感谢电子工业出版社的相关编辑，他们为本书的引进做出了贡献，同时对我们的翻译工作给予了大力支持。

在本书的翻译工作中，清华大学电子工程系语音与音频技术实验室的博士研究生和博士后也参与了部分内容的翻译工作，他们是（按姓氏拼音排序）：蔡猛、钱彦旻、单煜翔、史永哲、杨毅等，在此一并表示感谢。

本书虽然经过两次翻译校对，但是难免仍然会存在错误和不妥之处，欢迎读者批评指正。

刘加　张卫强　何亮　路程

2015 年 11 月于清华园

前　言

70 多年来，语音信号处理一直是一个活跃且不断发展的领域。最早的语音处理系统是模拟系统，如 20 世纪 30 年代由 Homer Dudley 及其同事们在贝尔实验室开发并于 1939 年在纽约世博会上展出的 Voder 系统，该系统可通过手工操作合成出语音；同期，Homer Dudley 在贝尔实验室还开发出了通道声码器或声音编码器；20 世纪 40 年代，Koenig 及其同事们在贝尔实验室开发出了声音语谱图系统，该系统可以在时域和频域展示语音的时变特征；另外，20 世纪 50 年代，全世界的很多研究实验室都开发出了早期的语音单词识别系统。

数字信号处理（DSP）起源于 20 世纪 60 年代，在 DSP 应用的广泛领域中，语音处理是其早期发展的驱动力。在此期间，先驱研究者们如麻省理工学院林肯实验室的 Ben Gold 和 Charlie Rader，贝尔实验室的 Jim Flanagan、Roger Golden 和 Jim Kaiser，他们开始研究数字滤波器的设计和应用方法，并用于语音处理系统的模拟仿真。随着 1965 年 Jim Cooley 和 John Tukey 发明快速傅里叶变换（FFT）技术以及 FFT 在快速卷积和谱分析方面的广泛应用，模拟技术的束缚和局限逐渐被打破，数字语音处理随之产生并展现出了清晰的面貌。

1968 年至 1974 年期间，本书作者（Lawrence R. Rabiner 和 Ronald W. Schafer）在贝尔实验室一起密切地工作，期间 DSP 领域取得了很多的基础性进展。当 Ronald W. Schafer 于 1975 年离开贝尔实验室并在佐治亚理工学院任学术职位时，数字语音处理领域已蓬勃发展，于是我们觉得是时候写一本关于语音信号数字处理方法和系统的教材了。到 1976 年，我们相信数字语音处理的理论发展得已经足够完备，精心撰写一本教材不但可以作为讲授数字语音处理基础知识的教材，还可以作为未来语音处理实际应用系统设计的参考书。1978 年，Prentice Hall 公司出版了这本教材《数字语音信号处理》。采用这本教材，Ronald W. Schafer 开设了第一门数字语音处理的研究生课程，期间 Lawrence R. Rabiner 仍在贝尔实验室从事数字语音处理基础的研究工作（Lawrence R. Rabiner 在贝尔实验室和 AT&T 实验室工作了 40 年，2002 年也进入学术界，在罗格斯大学和加州大学圣巴巴拉分校任教。Ronald W. Schafer 在佐治亚理工学院工作 30 年后，于 2004 年加入了惠普实验室）。

1978 年出版的教材的目标是，介绍语音基础知识和数字语音处理方法，以便构建强大的语音信号处理系统。从宏观层面来说，我们达到了最初的目标。本书按我们的预想服务了 30 多年，令我们高兴的是，直到今天它仍然广泛应用于本科生和研究生的语音信号处理课程教学。然而，根据我们过去 20 年来教授语音处理课程的经验，原书的基础尚可，但很多内容已与当代语音信号处理系统脱节，且未涉及当前的很多研究热点。这本新书正是我们改进这些问题的尝试。

在着手统一数字语音处理的现有理论和实践的艰巨任务时，我们发现原书中的很多内容还是正确且相关的，因此新书的起点很好。此外，我们从语音处理的科研和教学经验中了解到，1978 年出版的教材中，虽然内容组织基本上没有问题，但它已经不适合用来理解当代的语音处理系统。针对这些问题，我们在组织新书的内容时采用了新的框架，它与原书相比有两大改变。首先，我们包含了已有的数字语音处理知识体系结构。这种体系的第一层是语音基础科学和工程方面的基础知识；第二层是语音信号的各种表示。原书主要侧重了这两层，但一些关键主题则有所缺失。第三层是操作、处理和抽取语音信号中信息的各种算法，这些算法基于前两层的科学和技术知识。

顶层（即第四层）是语音处理算法的各种应用，以及处理语音通信系统中问题的技术。

我们努力按照这种体系结构（即语音金字塔）来展现新书的内容。为达到这一目的，第 2 章至第 5 章主要介绍金字塔的底层，内容包括语音产生和感知基础知识、DSP 基础知识回顾，以及声学、语音学、语言学、语音感知、声道中的声音传播等。第 6 章至第 9 章介绍如何通过基本的信号处理原理来表示数字语音信号（语音金字塔的第二层）。第 10 章介绍如何设计可靠和稳健的语音算法来估计感兴趣的语音参数（语音金字塔的第三层）。最后，第 11 章至第 14 章介绍如何利用语音金字塔前几层的知识来设计和实现各种语音应用（语音金字塔的第四层）。

新书在结构和行文上的一个重要变化是，为了尽可能地方便教学，我们在呈现内容时侧重于学习新思想的三个方面，即理论、概念和实现。对每个基本概念，我们都用很容易理解的 DSP 概念进行理论阐释；类似地，为了加深理解，每个新概念都提供了简单的数学解释和精心准备的例子与插图；最后，基于教学中对基础知识的理解，针对每个新概念的实现，提供了可实现特定语音处理操作的 MATLAB 代码（通常包含在每章中），每章的习题中配备了文档详尽的 MATLAB 练习。我们还在教学网站上提供了求解所有 MATLAB 练习所需要的内容，如 MATLAB 代码、数据库、语音文件等。最后，我们提供了几种语音处理系统结果的音频演示。通过这种方式，读者可以直观地了解各种语音信号处理后的语音质量。

更具体地讲，这本新书的组织如下。第 1 章简要介绍语音处理的领域，简要讨论贯穿于全书的主题的应用领域。第 2 章简要回顾 DSP 的概念，重点在于与语音处理系统密切相关的几个关键概念：

1. 从时域到频域的转换（通过离散时间傅里叶变换方法）。
2. 了解频域采样的影响（即时域混叠）。
3. 了解时域采样（包括下采样和上采样）的影响，以及频域的混叠和镜像。

在回顾 DSP 技术的基础知识后，第 3 章和第 4 章讨论语音的产生和感知。这两章与第 2 章和第 5 章一起，构成了语音金字塔的底层。从这里，我们开始讨论语音产生的声学理论，对不同的语音发音，我们导出了一系列声学语音模型，并展示了语言学和语音学如何与语音发声声学一起相互作用，生成语音信号及其在语言上的解释。讨论从语音在人耳中如何处理开始，到声音转换为通往大脑的听感知神经通路中的神经信号结束，我们通过分析语音感知过程，讨论了语音通信的基本过程，还简要讨论了几种在一些语音处理应用中可能嵌入语音感知知识到听感知模型的方法。第 5 章介绍关于人类声音在声道中传播问题的基础知识，表明与声道相似的均匀无损声管具有共振结构，以此阐明语音中的共振（共振峰）频率。还展示了如何通过适当的“终端模拟”数字系统来表示一系列级联声管的传播特性。该“终端模拟”数字系统具有特定的激励函数、对应不同长度和面积声管的特定系统响应，以及对应声音在唇端传输的特定辐射特征。

接下来的四章主要介绍 4 种数字语音信号的表示（语音金字塔的第二层）。第 6 章从语音产生的时域模型开始，逐步展示了如何通过简单的时域测量方法来估计模型中的基本时变属性。第 7 章介绍对语音信号应用短时傅里叶分析，以便实现无失真的分析/合成系统。取决于待处理信息的性质，我们解释了两种短时傅里叶分析/合成系统，两者都有着广泛的应用。第 8 章描述语音的同态（倒谱）表示，其中用到了卷积信号（如语音）可以转换为一系列加性分量这一性质。由于语音信号可以表示为激励信号和声道系统的卷积，因此语音信号非常适合于这种分析。第 9 章介绍线性预测分析的理论和实践，线性预测是语音信号的一种模型表示，当前的语音样本可以通过先前 p 个语音样本的线性组合建模表示，通过寻找最优线性预测器（最小均方误差）的系数，实现在给定时间段内最优的匹配语音信号。

第 10 章（语音金字塔的第三层）使用前面章节中介绍的信号处理表示和语音信号基础知识，

介绍了如何使用短时（对数）能量、短时过零率、短时自相关函数等测量值来估计基本的语音属性，例如分析的信号段是语音还是静音（背景信号）、语音段是浊音还是清音、浊音语音段的基音周期（基音频率）、语音段的共振峰（声道共振）等。对于许多语音属性，4 种语音表示中的每一种，都可以作为估计语音属性的高效算法使用。同时还介绍了如何基于 4 种语音表示中的两种测量法来估计共振峰。

第 11 章至第 14 章（语音金字塔的顶层）介绍语音和音频信号处理技术的几种主要应用。这些应用是深入理解语音和音频技术的成果。讨论语音应用的目的是，让读者基本了解如何构建这些应用，了解它们在不同比特率和不同应用场景下的性能。具体来讲，第 11 章介绍语音编码系统（包括开环和闭环系统）；第 12 章介绍如何使用感知掩蔽准则来构建具有最小编码感知误差的音频编码系统；第 13 章介绍如何构建口语对话系统中使用的文语转换合成系统；第 14 章介绍语音识别和自然语言处理系统，以及它们在一系列面向任务的场景中的应用。

本书可作为已先修 DSP 课程的学生的一个学期的语音处理教材。在我们自己的教学实践中，重点讲解第 3 章至第 11 章，同时选讲其他章节的部分内容，以便使学生对音频编码、语音合成和语音识别系统也有一定的认识。为了帮助教学，每章都提供了一些有代表性的课后习题，以强化每章讨论的概念。成功完成合理数量的课后习题，对理解语音处理的数学和理论概念非常重要。但如读者了解的那样，很多语音处理都是经验性的，因此我们提供了许多 MATLAB 练习来强化学生对语音处理基本概念的理解。我们还提供了配套的教学网站[①]（http://www.pearsonhighered.com/educator/product/Theory-and-Applications-of-Digital-Speech-Processing/9780136034285.page#dw_resources），并随时更新网站的内容，包括所需的语音文件、数据库和求解 MATLAB 练习的 MATLAB 代码，以及一系列语音处理概念的演示。

致谢

在语音处理的职业生涯中，我们非常幸运拥有过在杰出研究和学术机构的工作经历，这些单位为我们提供了充满激情的研究环境，并且鼓励我们分享知识。对于 Lawrence R. Rabiner 而言，这些单位包括贝尔实验室、AT&T 实验室、罗格斯大学和加州大学圣巴巴拉分校；对于 Ronald W. Schafer 而言，这些单位包括贝尔实验室、佐治亚理工大学 ECE 和惠普实验室。没有这些单位的同事和领导的支持与鼓励，这本书不会存在。

很多人对本书的内容有直接或间接的重大影响，但我们最应感谢的是 James L. Flanagan 博士，他是我们两人职业生涯中很多关键时期的导师和益友。Jim 为我们如何从事科研、如何清晰合理地呈现研究结果提供了指导。无论是对这本书还是对我们各自的职业，他的影响都是非常深远的。

感谢有幸合作并互相学习的其他人，包括我们的导师麻省理工学院的 Alan Oppenheim 教授和 Kenneth Stevens 教授，以及我们的同事佐治亚理工学院的 Tom Barnwell 教授、Mark Clements 教授、Chin Lee 教授、Fred Juang 教授、Jim McClellan 教授和 Russ Mersereau 教授。这些人既是我们的同事，又是我们的老师，我们感激他们的睿智和多年来的指导。

直接参与本书准备工作的同事包括 Bishnu Atal 博士、Victor Zue 教授、Jim Glass 教授和 Peter Noll 教授，他们都提供了见解深刻的成果，这些成果对本书中的很多内容产生了很大的影响。感谢其他人允许我们使用其发表物中的图表，包括 Alex Acero、Joe Campbell、Raymond Chen、Eric Cosatto、Rich Cox、Ron Crochiere、Thierry Dutoit、Oded Ghitza、Al Gorin、Hynek Hermansky、Nelson Kiang、Rich Lippman、Dick Lyon、Marion Macchi、John Makhoul、Mehryar Mohri、Joern Ostermann、David Pallett、Roberto Pieraccini、Tom Quatieri、Juergen Schroeter、Stephanie Seneff、Malcolm Slaney、Peter Vary 和 Vishu Viswanathan。

① 本书向授课教师提供英文原版教辅（习题解答、PPT），具体申请方式请参见书后的“教学支持说明”。

感谢朗讯-阿尔卡特公司、IEEE、美国声学学会和 House-Ear Institute 允许我们使用已发表或备档的图表。

同时要感谢 Prentice Hall 公司的那些帮助出版本书的人员，包括策划编辑 Andrew Gilfillan、责任编辑 Clare Romeo 和助理编辑 William Opaluch。还要感谢 TexTech International 公司负责文字编校工作的 Maheswari PonSaravanan。

最后，感谢赞助商 Suzanne Dorothy 对我们给予的关爱、耐心和支持。

Lawrence R. Rabiner 和 Ronald W. Schafer

目　录

第 1 章　数字语音处理介绍

本书内容包括从古老人类语言的研究到最新的计算机芯片。自贝尔创造性地发明电话以来，工程师和科学家就一直在研究语音通信，目的是发明更加高效的人与人之间及人与机器之间的通信系统。19 世纪 60 年代，数字信号处理（DSP）技术开始在语音通信研究中处于核心地位，今天 DSP 技术已让过去几十年的许多研究成果得到应用。期间，集成电路技术、DSP 算法和计算机体系结构方面的进展创造了很好的技术环境，为语音处理、图像和视频处理、雷达和声呐、医疗诊断系统及消费电子等领域提供了近乎无限的创新机会。重要的是，要注意到过去 50 多年里 DSP 技术和语音处理技术是共同发展进步的，语音处理技术的应用促进了 DSP 技术理论和算法研究的发展，而这些发展又在语音通信研究和技术领域得到了实际应用。我们有理由预计将来这种共生关系会一直持续下去。

为充分理解一门技术，譬如数字语音处理，我们必须进行三个层次的理解，即理论层次、概念层次和实践层次。图 1.1 描绘了这一技术金字塔[①]。就语音技术来说，理论层次包含以下方面：语音产生的声学理论，语音信号表示的基本数学知识，每种表示所关联语音的各种属性的推导，以及通过采样、混频、滤波等将语音信号和现实世界关联的信号处理的基本数学运算。概念层次涉及如何应用语音处理理论进行各种语音测量，以及估计和量化语音信号的各种属性。最后，为使一门技术充分发挥潜力，将理论层次和概念层次转换到实践层次必不可少，即能够实现语音处理系统来求解特殊的应用问题。这个过程涉及对某个应用的约束条件和目标的认识、工程上的取舍和判断、编写可工作的计算机代码（通常是用 MATLAB、C 或 C++编写的程序），或是运行在实时信号处理芯片（如 ASIC、FPGA、DSP 芯片）上的特定代码的能力。

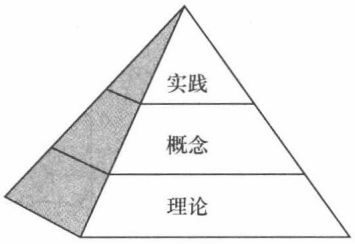


图 1.1　技术金字塔：理论、概念和实践

数字实现技术能力的持续改进反过来又开辟了新的应用领域，这些领域过去被认为是不可能或不现实的，这一道理同样适用于数字语音处理领域。因此，本书将重点放在对前面两个层次的理解上，但应时刻牢记最终的技术回报是在技术金字塔的第三层（实践层）。数字语音处理领域的基本原理和概念必然会继续发展和扩大，但过去 50 多年里学到的知识仍会为我们在未来几十年里看到的应用奠定基础。因此，对于本书涵盖的语音处理方面的各个主题，我们将努力使大家对理论层和概念层有尽可能多的理解；我们会提供一组练习题让读者在实践层次获得专业知识，这通常是通过每章后面习题中的 MATLAB 练习题来达到的。

在绪论的剩余部分，我们首先介绍语音通信过程和语音信号，最后介绍数字语音处理技术的重要应用领域。本书其他部分的设计是为了给读者对基本原理的学习打下扎实的基础，并强调 DSP 技术在现代语音通信研究和应用中扮演的核心角色。我们的目标是全面概述数字语音处理，内容涵盖了从语音信号的基本特性、以数字形式表示语音的各种方法，到话音通信及语音自动合成和识别应用的各个方面。在这个过程中，我们希望回答如下问题：

- 语音信号的本质是什么？
- 学习语音信号过程中 DSP 技术扮演怎样的角色？

---

① 使用术语“技术金字塔”而非“技术三角形”，是为了强调每层都有宽度和厚度并且支持更高的层。

- 语音信号有哪些基本的数字表示？它们在语音处理算法中怎样使用？
- 数字语音处理方法有哪些重要应用？

首先介绍语音信号以了解语音信号的性质。

1.1 语音信号

语音的基本目的是为了人类沟通，即说话者和倾听者之间消息的传输。据香农信息论[364]，以离散符号序列表示的消息可对其信息量以比特进行量化，信息传输速率可用比特/秒（bps）进行度量。在语音产生及许多人类设计的电子通信系统中，待传输信息以连续变化的波形（模拟波形）进行编码，这种波形可以传输、记录（存储）、操纵，最后被倾听者解码。消息的基本模拟形式是一种称为*语音信号*的声学波。如图 1.2 所示，语音信号可通过麦克风转换成电信号，进一步通过模拟和数字信号处理方法进行操纵，然后可根据需要通过扬声器、电话听筒或头戴式耳机转换回声学波。这种语音处理方式为贝尔发明电话奠定了基础，同时也是今天大多数记录、传输、操纵语音和音频信号的设备的基础。用贝尔自己的话说[47]："华生，如果我能得到一种像声音传播时空气改变密度那样改变电流密度的机制，就能通过电来传递任何声音，甚至是语音。"

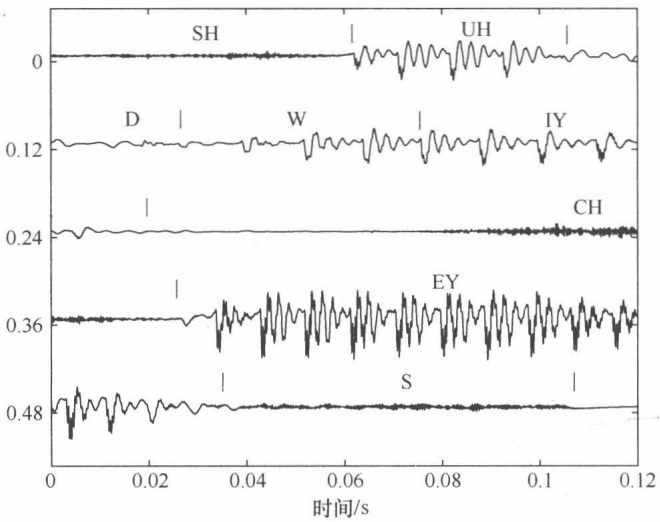


图 1.2 消息"should we chase"的语音波形，其中带有音素标记

虽然贝尔在不知道信息论的情况下有了伟大的发明，但信息论的原理在设计复杂的现代数字通信系统时发挥着巨大的作用。因此，尽管我们的重点在于语音波形和其参数化模型的表示，但讨论一下在语音波形编码中使用的信息论还是有用的。

图 1.3 形象化地展示了语音信号产生和感知的完整过程——从说话者大脑中消息的形成，到语音信号的产生，最后到倾听者对消息的理解。Denes and Pinson[88]在其语言学的经典介绍中，将这一过程称为*语音链*。图 1.4 给出了语音链的详细框图。这个过程从左上方开始，此时消息以某种方式出现在说话者的大脑中。在语音产生过程中，消息携带的信息可认为有着不同的表示形式（如图 1.4 上面的路径所示）。例如，消息最初可能以英语文本的形式表示。为"说出"这条消息，说话者隐式地将文本转换成对应口语形式声音序列的符号表示。该步骤在图 1.4 中被称为*语言码生成过程*，它将文本符号转换成音素符号（伴随着重音和段长信息），音素符号用来描述口语形式消息的基本声音及声音产生的方式（即语速和语调）。例如，若将图 1.2 中的波形片段用一种便

于计算机键盘输入的 ARPAbet[①]代码标记，则文本“should we chase”按照发音可表示成[SH UH D - W IY - CH EY S]（关于音素标注的详细讨论，见第 3 章）。语音产生过程的第三步是转变成“神经肌肉控制”，这组控制信号指引神经肌肉系统以一种与产生口语形式消息及其语调相一致的方式，移动舌头、唇、牙齿、颌、软腭这些发音器官，神经肌肉控制这一步的最终结果是产生一组关节运动（连续控制），使声道发音器官按照规定的方式移动，进而发出期望的声音。语音产生过程的最后一步是“声道系统”，声道系统产生物理声源和恰当的时变声道形状，产生图 1.2 所示的声学波形。通过这种方式，期望表达消息中的信息就被编码为语音信号。

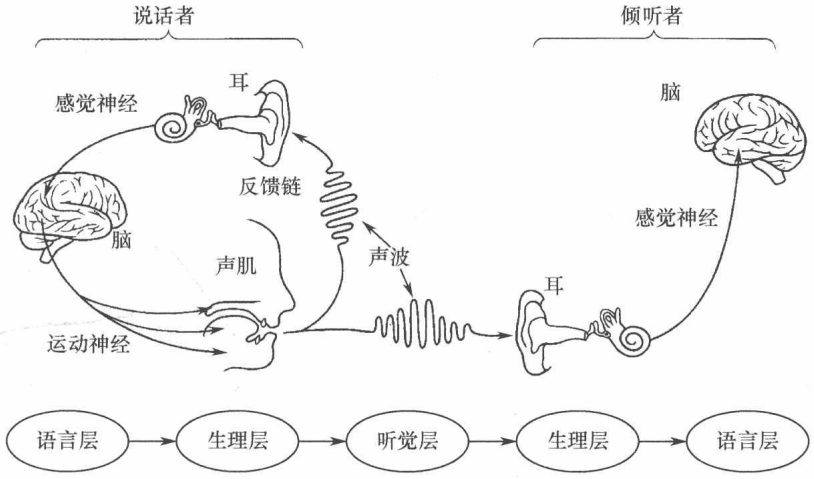


图 1.3　语音链：从消息到语音信号再到理解（据 Denis and Pinson[88]）

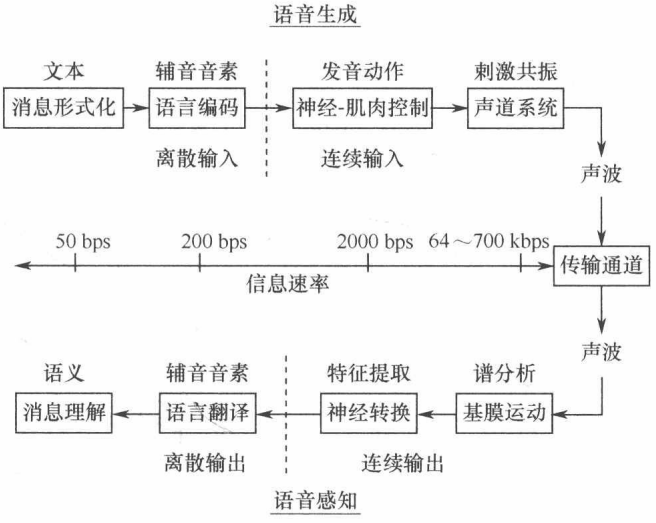


图 1.4　语音链的框图表示

为了确定语音产生过程中信息流的速率，我们假设在书面语中约有 32 个符号（字母，英语中有 26 个字母，若包括标点符号和空格，则接近 $32 = 2^5$ 个符号）。正常的平均说话速率约为 15 个符号每秒，因此，假设字母相互独立后做简单的一阶近似，文本消息编码成语音后的基本信息速率约为 75bps（5 比特每符号乘以 15 个符号每秒）。但是，实际的速率会随着说话的速率变化而变化。

① 国际语音协会（IPA）为音素标注提供了一套规则，它用等价的一组特殊符号来表示音标。ARPAbet 编码不需要特殊字体，因此更加便于计算机应用。

对于图 1.2 中的例子，文本包含 15 个字母（包括空格），对应的语音词条持续了 0.6 秒，因此有更高的速率 15×5/0.6 = 125bps。在语音产生过程的第二个阶段，文本表示转变成基本声音的单元，它们称为带有韵律（即音高和重音）标记的音素，此时信息速率很容易达到 200bps 以上。图 1.2 中用来标注语声片段的 ARBAbet 音素集包含近 64 = 2^6 个符号，即 6 比特每音素（假设音素相互独立得到的粗略近似）。在图 1.2 中，大约 0.6 秒的时间里有 8 个音素，计算得到信息速率为 8×6/0.6 = 80bps，考虑描述信号韵律特征的额外信息（如段长、音高、响度），文本信息编码成语音信号后，总信息速率需要再加上 100bps。

语音链前两个阶段的信息表示是离散的，所以用一些简单假设就可估计信息流的速率。在语音链中语音产生部分的下一阶段，信息表示变成连续的（以关节运动时的神经肌肉控制信号的形式）。若它们能被度量，就可估计这些控制信号的频谱带宽，进行恰当的采样和量化获得等效的数字信号，进而估计数据的速率。与产生的声学波形的时间变化相比，关节的运动相当缓慢。带宽估计和信号表示需要达到的精度要求意味着被采样的关节控制信号的总数据率约为 2000bps[105]。因此，用一组连续变化信号表示的原始文本消息传输，比用离散文本信号表示的消息传输需要更高的数据率①。在语音链中语音产生部分的最后阶段，数字语音波形的数据率可从 64000bps 变化到超过 700000bps。我们是通过测量表示语音信号时为达到想要的感知保真度所需要的采样率和量化率计算得到上面的结果的。例如，“电话质量”的语音处理需要保证带宽为 0～4kHz，这意味着采样率为 8000 个样本/秒。每个样本可以用对数尺度量化成 8 比特，从而得到数据率 64000bps。这种表示方式很容易听懂（即人们可很容易地从其中提取出消息），但对于大多数倾听者来说，语音听起来与说话者发出的原始语音会有不同。另一方面，语音波形可以表示成“CD 质量”，即采用 44100 个样本/秒的采样率，每个样本 16 比特，总数据率为 705600bps，此时复原的声学波听起来和原始语音信号几乎没有区别。

当我们通过语音链将文本表示变成语音波形表示时，消息编码后能够以声学波形的形式进行传播，并且可被倾听者的听觉机制稳健地解码。前面对数据率的分析表明，当我们将消息从文本表示转换成采样的语音波形时，数据率会增大 10000 倍。这些额外信息的一部分能够代表说话者的一些特征，如情绪状态、说话的习惯、口音等，但主要是由简单采样和对模拟信号进行精细量化的低效性导致的。因此，出于语音信号固有的低信息速率的考虑，很多数字语音处理的重点是用比采样波形更低的数据率对语音进行数字表示。

完整的语音链包括上面讨论的语音产生/生成模型，也包括图 1.4 底部从右向左显示的语音感知/识别模型。语音感知模型显示了从耳朵捕捉语音信号到理解语音信号编码中携带的消息的一系列处理步骤。第一步是将声学波有效地转换成频谱表示，这是由耳朵内部的基底膜实现的，基底膜的作用类似于非均匀频谱分析仪，它能将输入语音信号的频谱成分进行空间分离，以非均匀滤波器组的方式进行频谱分析。语音感知过程中的第二步是神经传导过程，将频谱特征变成可被大脑解码和处理的声音特征（或语音学领域中所指的差异性特征）。第三步通过人脑的语言翻译过程将声音特征变成与输入消息对应的一组音素、词和句子。语音感知模型中的最后一步是将消息对应的音素、词和句子变成对基本信息意义的理解，进而做出响应或采取适当的处理。我们对图 1.4 中大部分语音感知模块过程的基本理解还是非常初步的，但人们普遍认为语音感知模型中各个步骤物理间的相互关联发生在人脑中，因此整个模型对于思考语音感知模型中各个过程的发生非常有帮助。第 4 章中将讨论听觉和感知机理。

图 1.4 所示的整个语音链框图中还有一个过程我们没有讨论，即模型中语音产生部分和语音感

① 为数字表示引入术语“数据率”，是为了区别于语音信号表示的消息中所含的内在信息内容。