语音信号处理与识别

Speech Processing and Recognition

严勤 吕勇 著

国防工业出版社
National Defense Industry Press

语音信号处理与识别

严勤　吕勇　著

国防工业出版社

·北京·

内 容 简 介

本书系统介绍语音信号处理的理论、方法和应用，着重讨论英语口音的分析与转换、语音增强和鲁棒语音识别。全书共分10章，内容包括语音信号处理概述、语音信号模型及声学特征、鲁棒语音识别的基本方法、英语口音的声学差异、英语口音的声学分析、英语口音转换、基于共振峰曲线和谐波噪声模型的语音增强、基于特征补偿的鲁棒语音识别、基于矢量泰勒级数的多环境模型自适应算法和基于多项式回归的模型自适应算法。

本书可供信息与通信工程、计算机科学与技术等专业的教师、科研人员以及研究生使用。

图书在版编目(CIP)数据

语音信号处理与识别/严勤，吕勇著. —北京：国防工业出版社，2015.12

ISBN 978-7-118-10583-4

Ⅰ.①语… Ⅱ.①严… ②吕… Ⅲ.①语声信号处理 Ⅳ.①TN912.3

中国版本图书馆CIP数据核字(2015)第305613号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路23号 邮政编码100048)

三河市众誉天成印务有限公司印刷

新华书店经售

*

开本 710×1000 1/16 印张 12¼ 字数 250千字

2015年12月第1版第1次印刷 印数 1—2000册 定价 62.00元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

前　言

PREFACE

语音信号处理是以语音学和数字信号处理为基础的综合性技术，涉及声学、计算机、通信、模式识别、人工智能、心理学、生理学等多个学科。语音信号处理的应用领域主要包括语音增强、语音编码、语音合成和语音识别。目前，在实验室理想环境下，各领域的研究已经取得了较好的成果。然而，在实际应用中，语音处理系统会工作于各种复杂的环境中，实际环境与理想环境的不匹配会导致系统的性能下降。因此，研究语音信号的变异性，提高系统的环境鲁棒性，成为语音处理系统走向实际应用的关键。

本书在系统介绍语音信号处理基本理论的同时，着重讨论语音处理的口音分析与转换、基于共振峰曲线与谐波噪声模型的语音增强和噪声环境下语音识别系统的鲁棒性，这些内容致力于减小语音本身的变异性和外部环境噪声对语音信号处理的影响，提高语音处理系统的鲁棒性。

本书共分10章，内容分为四个部分：语音信号处理的基础知识（第1～3章）、口音的分析与转换（第4～6章）、基于共振峰曲线和谐波噪声模型的语音增强（第7章）和鲁棒语音识别研究（第8～10章）。

第1章主要介绍语音的基础知识、语音口音的研究现状和语音识别的基本原理。第2章介绍语音信号处理的基本模型和声学特征。语音模型包括源—滤波器模型、线性预测模型和贝叶斯模型，在贝叶斯模型中着重讨论隐马尔可夫模型。声学特征包括共振峰、美尔频率倒谱系数、语调特征与语速特征。第3章介绍鲁棒语音识别的基本方法，包括特征域方法和模型域方法。前3章属于语音信号处理与识别的基本理论，它们是后续章节研究内容的基础。

第4章介绍英语口音的声学差异，包括英语口音的演化与分布、语音学特点和声学特点。第5章对英语口音进行了声学分析，包括共振峰及共振峰轨迹的概率模型、英语口音的共振峰特征分析和英语口音的韵律分析。第6章研究英语的口音转换，包括共振峰转换、语调转换和并行口音转换。语音的口音分析与转换研究可以减小语音本身变异性对语音处理系统的影响，提

高系统的鲁棒性。

第7章研究基于共振峰曲线和谐波噪声模型的语音增强，包括噪声环境下共振峰曲线提取、谐波噪声模型及两者基础上的语音增强。与最小均方误差方法相比，该算法由于保持了原信号的时频结构完整性，因而可以显著降低由信号处理引起的噪声。

第8章研究了基于特征补偿的鲁棒语音识别，包括基于隐马尔可夫模型的模型组合和基于矢量泰勒级数的自适应特征补偿，这些研究致力于在特征域减小外部环境噪声对语音识别系统的影响，从含噪语音中提取纯净语音信息。第9章将模型自适应算法从单环境模型（纯净语音模型）扩展到多环境模型，用多个基本训练环境声学模型预测实际测试环境，提高模型自适应的精度。第10章研究基于多项式回归的模型自适应和基于子带多项式回归的模型自适应，它们属于基于变换的间接模型自适应，可以克服任意语音变异性的影响，适合多种环境失配因素存在时的鲁棒语音识别。

作者在河海大学计算机与信息学院从事信号处理与模式识别方面的研究工作，本书是作者在语音信号处理领域近年来科研成果基础上的总结性著作，理论阐述详细，叙述条理清晰，具有充分的实验数据支持，适合有志于该领域研究的相关人员参考使用。作者在本书中述及的相关研究工作得到国家自然科学基金（61170297，61301218）和江苏省自然科学基金（BK2010520）的资助。感谢河海大学在本书出版过程中提供的帮助与支持。硕士研究生邓舒宇、印晶晶、何晓梅参加了相关材料的整理工作，在此表示衷心感谢！

由于作者水平有限，书中难免存在一些不足和不妥之处，诚恳希望广大读者批评指正。

作　者

2015年7月于南京

目　录

CONTENTS

第1章

语音信号处理概述

语音是人的发声器官发出的具有一定语言意义的声音，是人类交流信息的基本手段，是最自然、最有效、最方便的人际交互工具，是语言的声学表现形式。语音学是语言学的一个分支，主要研究语音的发音机制、语音的声学特性和语音的听觉感知过程。语音信号处理是以语音学和数字信号处理为基础的综合性学科，涉及声学、计算机、通信、模式识别、人工智能、心理学、生理学等多个学科。

从应用方面划分，语音信号处理研究主要包括语音增强、语音编码、语音合成、语音识别和说话人识别等分支。语音增强是语音信号处理的基础，是消除语音通信中噪声干扰的有效方法，目的是从背景噪声中提取、增强有用语音信号，抑制、降低噪声干扰，尽可能从带噪语音信号中提取纯净的原始语音。实际生活中，语音不可避免地要受周围环境背景噪声、传输系统内部噪声甚至其他说话者的干扰。噪声会降低语音的可懂度，影响接收者的情绪，很强的噪声甚至会完全掩蔽语音，使语音变得完全不可懂。因此，在语音编码、语音识别、语音合成等语音处理系统的前端有必要采取语音增强技术抑制环境噪声，提高语音质量。

语音编码是语音通信和语音存储的基础，目的是对数字语音进行压缩，以提高语音的传输效率，减小语音占用的存储空间。相对于模拟语音，数字语音具有抗干扰能力强、易于复制和保存等优点，但其占用空间大。例如，1min 采用 44.1kHz 采样率，16 位采样精度的双声道语音信号占用的存储空间高达 10MB。如此大的数据量给语音的存储和传输带来了极大挑战。实际上，数字语音中含有大量的冗余信息，通过各种语音编码技术去除语音信号的冗余度，就可以达到对语音压缩的目的。

语音合成和语音识别是实现人机语音通信的两项关键技术，前者将文本转换为语音，后者将语音转换为文本。语音合成是一项比较成熟的语音处理技术，目的是让机器“说话”，将以其他形式存储的信息转换为语音。文语转换技术是语音合成的一个重要分支，目的是将文字智能地转化为自然语音流。文语转换

是一项比较成熟的语音处理技术，合成的语音质量较高，已经广泛应用于自动报站、电话查询、语音玩具等领域。

让机器“听懂”人类口述的语言，与机器进行语音交流，一直是人类追求的目标。语音识别的研究目的是将人类的口语信号转变为相应的文本或指令。它包括两个方面的含义：一是将人类口述的语言逐词逐句转换为相应的文本，即书面语言；二是将口语中的命令或要求提取出来，使机器能够接收人的口语指令，理解人的意图，从而做出正确的响应。与语音合成相比，语音识别系统的实现难度较大，虽然在实验室理想环境中可以取得很高的识别率，但是在实际应用中极易受环境噪声的影响，识别性能急剧下降。因此，语音识别技术的商品化还存在许多待解决的问题。

说话人识别与语音识别类似，都是通过提取语音信号的特征参数，根据训练阶段得到的声学模型对特征参数进行判别，判断其属于哪一类。它们的区别主要在于识别目的不同：语音识别的目的是提取不同说话人同一词语发音的共性，即提取语音的语义信息；说话人识别的目的是提取语音信号中不同说话人之间的个性特征，即不同说话人之间的特征差异。说话人识别又分为说话人确认和说话人辨认：前者只要判断当前发音是否属于某个说话人，只需做出“是”或“否”两种判断，可以取得很高的识别率；而后者需要判断当前发音属于若干个候选说话人中的某一个，其实现难度较大。

1.1 语音基础知识

1.1.1 语音的声学特性

语音是一种特殊的声音，即人类说话的声音，是语音信息的声学表现。声波是一种由机械振动引起的纵波，可以通过气体、液体或固体等传输介质传播，并能够被人或动物的听觉器官所感知。下面介绍描述声波的一些基本物理量。

1. 频率与速度

声波的频率由声源决定，定义为声源每秒振动的次数，单位为赫兹(Hz)。人耳所能听到的声波频率为 20Hz ~ 20kHz，低于 20Hz 的声波称为次声，高于 20kHz 的声波称为超声。语音的频率为 300Hz ~ 3400Hz，高于 3400Hz 的频率成分对语音的可懂度和自然度影响较小，通常忽略。

实验表明，人耳对声音频率高低的感觉与实际频率的高低不呈线性关系，而是近似呈对数关系，因此，在语音声学中常用对数频率或美尔(Mel)频率等非线性频率刻度。周期是声源振动一次所需要的时间，即频率的倒数，单位为秒(s)。

声波的传播速度由传输介质决定，在不同介质中传播时，固体中的速度最快，液体次之，气体最慢。传播速度不仅与介质的种类有关，还与介质的温度有关。例如，在15℃时，空气中的声波速度为340m/s。声波是一种纵波，其传播方向与振动方向一致，声波在一个周期内传播的距离称为波长。因此，波长由声波的频率和速度共同决定。波长可以表示为

$$\lambda = cT = \frac{c}{f} \tag{1-1}$$

式中：f 为频率；T 为周期；c 为声速。

2. 声压与声压级

当声源在空气中振动时，会压缩空气，导致大气压强发生变化，因此常用声压描述声波。其定义为实际气压 p' 与静态气压 p_0 的差值，即

$$p = p' - p_0 \tag{1-2}$$

声压 p 的单位与气压相同，都是帕斯卡（Pa）。在声场中，每一点的声压都是瞬时变化的，因此常用其有效值来表示。声压的有效值定义为一段时间内各个采样时刻的瞬时声压的均方根，即

$$P_{\text{rms}} = \sqrt{\frac{p_1^2 + p_2^2 + \cdots + p_n^2}{n}} \tag{1-3}$$

式中：p_1、p_2、…、p_n 为各个采样时刻的瞬时声压。

人耳对声波强弱的感觉不与声压的大小成正比，而是与声压的对数近似呈线性关系，因此常用分贝（dB）为单位表示声音的相对强弱。人耳所能感知到的最小声压（闻阈）$P_0 = 2 \times 10^{-5}$Pa，将其定义为0dB。有效声压的声压级定义为

$$L = 20 \lg \frac{P}{P_0} \tag{1-4}$$

人耳所能承受的最大声压（痛阈）约为20Pa，对应的声压级为120dB，安静的房间中的声压级为20～30dB，城市道路上的声压级为60～80dB。如果声压级超过100dB，人就会感到很不舒服。

3. 纯音与复音

纯音是指单一正弦振动引起的声波，复音是指包含多个正弦声波的声音。自然界中纯音很少，大部分是复音。在复音中，各个正弦波频率的最大公因数称为基音频率。基音频率所对应的正弦波成分就称为基音。频率等于基音频率正整数倍的正弦波成分都称为谐音（谐波）。

语音的基音频率为60～500Hz，女声的基音频率比通常男声的高一些。语音的最高谐波频率可达5000Hz以上，通常包括几十次到几百次的谐波成分。

4. 共振与共振峰

当物体的振动频率等于其固有频率时，就会发生共振。作为由振动引起的机械波也会发生共振现象。在人的发音器官中，声带属于声源，咽腔、口腔、鼻腔可看作调制系统，声带振动发出的声音，经过咽腔、口腔、鼻腔调制传播到空气中。由咽腔、口腔、鼻腔等组成的调制系统可以用线性滤波器来近似描述，由于该滤波器存在多个极点，因此语音经过其调制后会具有明显的共振峰结构。

1.1.2 语音的基本单元

音素是语音的最小单位，一般分为元音和辅音两种。当口腔、唇腔等完全开放时，声带振动发出的声音气流可以顺利通过，这时产生的语音称为元音。当发声通路不完全开放时，声带发出的声音气流在发声器官中就会受到阻碍，这时产生的语音称为辅音。

任何一种语言都有若干个不同的元音，决定元音音色的主要因素是舌头形状及其在口腔中的位置与嘴唇的形状。从声学特性来看，元音具有明显的共振峰结构，当元音进入声道（咽腔、口腔、鼻腔）时会引发共振，产生一组共振频率，称为共振峰频率，简称共振峰。共振峰是区别不同元音的重要参数，一个元音的共振峰参数包括每个共振峰频率值及其频带宽度，不同元音具有不同的共振峰结构。在实际应用中，一般只需用前三个共振峰即可区别不同的元音。

辅音没有明确的共振峰结构。辅音发音时的阻碍位置称为调音点。阻碍方式称为调音方式。根据调音方式的不同，辅音分为塞音、摩擦音、鼻音等多种类型。根据发音过程中声带是否振动，可以将辅音分为浊辅音和清辅音，声带振动的是浊辅音，声带不振动的是清辅音。

在语音学中，将声带振动发出的语音称为浊音，声带不振动发出的语音称为清音，浊音包括元音和浊辅音。浊音的声带振动频率称为基音频率，一般用 F_0 表示。在一段连续语音中，浊音段的 F_0 是随时间变化的，F_0 的变化就产生了声调。F_0 的变化轨迹称为声调轨迹。声调反映了语音的韵律特性，在汉语等语言中有辨意作用。

在语言中，具有一个响亮中心，可以被人的听觉器官明显感知的语言片断称为音节。一个音节可以由一个或多个音素组成。在汉语中，一个汉字的读音就是一个音节，普通话无调音节有 400 多个，有调音节有 1300 多个。在英语中，一个元音音素可以构成一个音节，也可以由一个元音因素和几个辅音因素结合构成一个音节。一个英语单词可以由一个、两个或多个音节组成，分别称为单音节单词、双音节单词和多音节单词。英语的音节总数量非常庞大，因此在语音识别

等应用中无法以音节为基本语音单元进行声学建模。

元音是一个音节的核心，在发音长度和能量方面都占据音节的主要部分。音节中的响亮中心即为元音的发音。辅音则出现在音节的前端、后端或前后两端，它们的时长和能量与元音相比都很小。音节具有显著的开始、中心、结尾三部分结构，因而音节之间可以明显地被感知，一段连续语音可以划分为若干个音节。

1.2 英语口音处理

1.2.1 英语口音概述

英语由古代从北欧及斯堪的纳维亚半岛移民至不列颠群岛的盎格鲁、撒克逊和朱特部落的日耳曼人所说的语言演变而来。由于近代英国的殖民活动，英语传播到了世界各地，被多个国家作为官方语言或第二语言。但是英语在不同国家及同一国家的不同地区的发音方式也有所不同，这就产生了英语的不同口音。英语口音的变化分为跨口音变化和口音内变化。跨口音变化是指不同国家英语标准口音间的变化，如英式口音、美式口音、澳大利亚口音、爱尔兰口音和苏格兰口音，它们各自有明显的标识性区别。某种口音的口音内变化主要指其地区性的口音变化。

与我国的普通话类似，以英语为官方语言的国家或地区通常会规定一种口音作为本国或本地区的“标准”发音。在英格兰，尽管没有法定的官方口音，但英式标准发音（Received Pronunciation，RP）是公认的标准发音。RP 是 11 世纪时形成于英格兰中南部的一支方言，该地区包含牛津和剑桥两个大学城。14 世纪时，RP 在贸易商人中广泛使用。同时，由于牛津大学和剑桥大学的崛起，RP 成为受过良好教育人士的语言。在 19 世纪到 20 世纪，RP 成为英国公立学校的教学语言，被英国广播公司（BBC）的播音员所使用，因此 RP 又称为 Public School English 和 BBC English。RP 是一种标准正统的中性英语口音，因而广泛用于非英语国家的英语教学中，非英语国家学习的英语一般都是 RP。RP 的地位虽然重要，但是在英国的日常生活中讲 RP 的人口比例非常小，各地区都有自己的“方言”，其发音方式或大或小有差别。例如：在伦敦地区，伦敦腔是一种流行的口音；在英格兰中部以伯明翰为中心的地区，鼻音很重，这种口音叫做 Brummie；在苏格兰地区，英语的多个元音有变异，这种口音叫做 Jock。

在美国，没有一种发音可以和英式标准发音（RP）相对应。美式标准发音仅是 2/3 美国人口音的一种通用称呼，这部分人的口音没有显著的地域性差别。

而且美式发音也有一定的口音内变化。

在澳大利亚，英语口音更接近英式发音。但是由于外来大众媒体产品在澳大利亚占据垄断地位，澳大利亚人也习惯使用一些美式英语的特有词汇和习惯用语。相对于英国，澳大利亚版图辽阔，但其英语发音区别较小和非常同质化。与英式发音的地域性特点相比，澳大利亚发音有乡村口音和城市口音的区别。此外，澳大利亚口音的变化更多的是社会阶层和语体性质的变化，而不是地域的变化。

1.2.2 英语口音的研究现状

由于不同人群的口音差别，语音处理技术必须充分考虑口音差异引起的发音变异。在语音识别中，测试说话人与训练说话人之间的口音失配会导致单词误识率上升30%左右。在语音合成中，数据库通常采用单一说话人的语音，因此合成语音被固定为同一种口音。在口音描述和识别中，一般有三种基本方法对由口音导致的发音变异进行建模。

1. 全局声学分布建模

对口音群体进行描述最简单的方法是：对该口音群体说话人的特征矢量的概率分布进行建模，得到声学模型。例如，可以用高斯混合模型对同一地区人群发音的谱包络概率分布进行建模，然后进行与发音内容无关的口音识别。全局声学分布建模实现简单，但是处理精度有限。尤其当所建模型包含其他语音变异性时，口音识别的结果不是很可靠。

2. 特定口音的音素模型

如果已知每种口音说话人的发音文本，就可以为每种口音建立一个音素模型集。在未知发音测试中，根据该模型集进行计算并选出具有最高概率的音素集，从而进行口音识别。这种方法的缺点是：由于识别器不一定会使用相同的最优音素序列，音素变异没有被充分利用。

3. 说话人发音系统分析

当口音群体数量较少时，基于特定口音音素模型的口音识别系统能较好地完成口音识别任务。然而，这种技术在处理一种语言的大量地区口音时性能并不明确。在说话人发音系统研究中，首先对单一说话人的音素模型集进行聚类分析，然后通过音素树建立说话人的发音特性，从而识别说话人的发音与标准发音的差别。类似地，在口音描述和识别中，可以使用语音片断之间的相似性描述某说话人的发音系统特点，并与已知口音的一般说话人发音系统进行比较以便识别口音。

1.3 语音识别

1.3.1 语音识别的基本原理

根据识别对象的不同，语音识别系统可分为孤立词识别、连接词识别和连续语音识别三种。随着语音识别技术的发展，语音识别的实现方法也随之变化，这里主要介绍基于倒谱特征参数和概率模型的统计语音识别原理。

图 1－1 为连续语音识别系统的原理框图，主要由特征提取、声学模型、声学解码、词表匹配、语言模型和语言解码等模块组成。

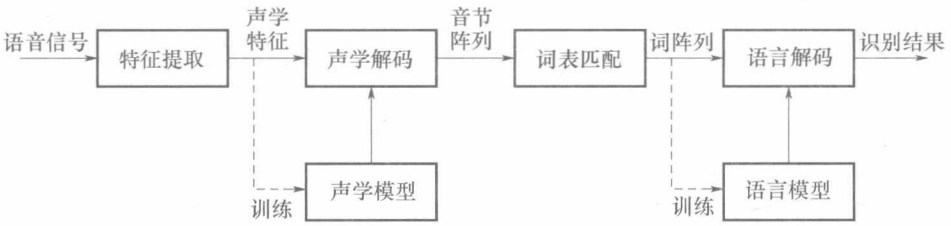

图 1－1 连续语音识别系统的原理框图

在特征提取模块，语音信号经过数字化、分帧、加窗、离散傅里叶变换等一系列处理，生成特征矢量序列。语音是一种非平稳的随机信号，但是在较短时间段内（10～30ms），可认为人的声带振动和声道形状保持不变。这段时间内的语音信号是平稳的，所以语音信号是一种短时平稳的随机信号。因此，窗函数的长度（帧长）一般取 10～30ms，将窗函数在时间轴上移动，即可将一段连续语音信号分割成若干帧。为保持帧信号之间的平滑性，前一帧与后一帧之间会有部分采样点重叠，因此帧移一般小于帧长，通常取帧长的 50%。得到帧信号后，对每一帧进行离散傅里叶变换得到功率谱或幅度谱，再进行相应的变换得到每一帧的特征矢量。目前，常用的语音识别系统多以美尔频率倒谱系数（Mel－Frequency Cepstral Coefficients，MFCC）为特征参数。如果输入语音是训练语音，则输出的声学特征用于训练，生成声学模型。如果输入语音为测试语音，则用声学模型对特征矢量序列进行声学解码，得到音节阵列。

声学模型用于描述每个基本语音单元特征矢量的概率分布，在训练阶段由语音库中每个基本语音单元的多个说话人的训练语音特征矢量训练生成。目前，多数语音识别系统用隐马尔可夫模型（Hidden Markov Model，HMM）对基本语音单元进行声学建模。对于汉语语音识别系统，可以用音节作为基本语言单