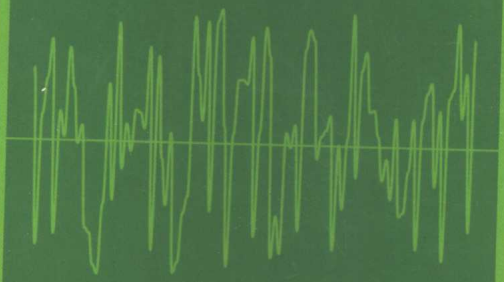
数字信号处理参考教材系列

语音与图像的数字信号处理

〔日〕谷萩隆嗣　编著



科学出版社
www.sciencep.com

数字信号处理参考教材系列

语音与图像的数字信号处理

〔日〕谷萩隆嗣　编著
朱　虹　译

科学出版社
北京

图字:01-2003-1057 号

ディジタル信号処理ライブラリー3
音声と画像のディジタル信号処理
Digital Signal Processing of Speech and Images
谷萩隆嗣 株式会社コロナ社
Takashi Yahagi CORONA PUBLISHING CO., LTD.

图书在版编目(CIP)数据

语音与图像的数字信号处理/(日)谷萩隆嗣编著,朱虹译.—北京:科学出版社,2003
(数字信号处理参考教材系列)
ISBN 7-03-011466-3

Ⅰ.语… Ⅱ.①谷… ②朱… Ⅲ.①语音信号处理 ②图像处理-数字技术 Ⅳ.①TN912.3 ②TN911.73

中国版本图书馆 CIP 数据核字(2003)第 040823 号

责任编辑 王 炜 崔炳哲 **责任制作** 魏 谨
责任印制 刘士平 **封面设计** 李 力

科学出版社 出版
北京东黄城根北街 16 号 邮政编码:100717
http://www.sciencep.com
源海印刷有限责任公司 印刷
北京东方科龙图文有限公司 制作
http://www.okbook.com.cn
科学出版社发行 各地新华书店经销

2003 年 9 月第 一 版 开本:A5(890×1240)
2003 年 9 月第一次印刷 印张:8 5/8
印数:1—4 000 字数:231 000

定 价:22.00 元
(如有印装质量问题,我社负责调换(新欣))

“数字信号处理参考教材系列”序

近年来，随着数字技术的惊人发展，以前用模拟技术进行处理或者以往根本无法进行数字处理的问题，都可以进行数字处理了。因此，数字技术越来越广泛地应用于诸多领域，而且这些领域对数字技术的要求也变得越来越高。

最近对电气、电子、信息、通信等领域进行的大规模市场调查表明，很多企业以及研究机构都对数字信号处理技术非常重视，他们在调查问卷的表格中，把数字信号处理填在了“必要性”和“重要性”一栏的首位。从这一社会现象也可以看出，数字信号处理是当今社会急需发展的学科领域之一。

鉴于这种状况，我们以供从事数字信号处理或者准备学习数字信号处理的社会各界人士参考阅读为目的，从更广泛的角度对数字信号处理这一学科进行归纳整理，编写了这套系列书。

本系列书包括以下各册：

1. 数字信号处理基础理论
2. 数字滤波器与信号处理
3. 语音与图像的数字信号处理
4. 快速算法与并行信号处理
5. 卡尔曼滤波器与自适应信号处理
6. ARMA 系统与数字信号处理
7. VLSI 与数字信号处理
8. 信息通信与数字信号处理

9. 人工神经网络与模糊信号处理

10. 多媒体与数字信号处理

上述各册中，第 1 至第 3 为基础部分，以大学三、四年级本科生为读者对象；第 4 至第 6 为比基础部分内容较深的提高部分，以研究生或者具有同等学历的科研人员及技术人员为读者对象；第 7 至第 10 为应用部分，以大学或研究机构的研究人员为主要读者对象，亦可供有一定基础知识的社会各界人士参考阅读。

也就是说，读者可根据自己的兴趣和所掌握的知识基础，有选择地阅读本系列书中的内容。比如，从基础知识开始学习数字信号处理的读者，可选择基础部分的内容；如果已具备了一定的基础知识，则可选择提高部分或者应用部分。从基础知识开始学习的，可按基础部分→提高部分→应用部分的顺序，或者按基础部分→应用部分→提高部分的顺序，根据自己的兴趣有选择地阅读。

本系列的执笔者均为目前仍活跃在相关领域第一线的专家、学者，因而编者有理由相信本系列书能够满足不同层次读者的需求。

另外，考虑到数字信号处理理论及应用技术的迅速发展，今后我们会根据情况及时补充新内容，使本系列书不断充实和完善。

最后，时值本系列书出版之际，谨向对本系列书的出版提供多方帮助的 CORONA 社的各位表示衷心的感谢。

“数字信号处理参考教材系列”策划兼主编

谷萩隆嗣

前　言

语音和图像是与听觉和视觉相关的最贴近我们日常生活的信息。随着数字信号处理技术的飞速发展，语音和图像中的重要部分和复杂部分几乎都可以用数字信号处理手段来处理。

本书首先分析语音的数字信号处理，对其基础、语音分析合成以及语音识别中的几种具有代表性的方法进行详细的论述；其次，以图像为研究对象，对图像数据处理时的基础知识和图像的增强、恢复及压缩等进行介绍。进一步对如CT图像处理中所需要的由投影图像来重构原始图像的方法进行详细的论述。

第1章，为了让读者掌握语音信号处理的概要，介绍基于语音的信息传递、人类语音的生成及对语音的感知、语音的生成模式、语音的合成与识别等与语音信号处理相关的基础知识。

第2章，详细介绍在语音分析合成系统中具有代表性的线性预测法。首先介绍线性预测的基本概念及线性预测法，并介绍对线性预测的缺陷进行改进后的PARCOR法和LSP法，以及各个方法的特点。

第3章，介绍与线性预测法并立的语音分析合成方法中的倒谱法。首先介绍倒谱法的概念，接下来介绍对倒谱法的缺陷进行改进后的改进型倒谱法和无偏倒谱法。

第4章，介绍语音信号处理中重要的语音规则合成及语

音识别的具有代表性的几种方法。介绍语音的规则合成及其概要，介绍语音识别中的具有代表性的 DP 匹配法和 HMM 模型法，论述连续语音的识别以及语音识别系统的非指定语者的对应方法。

第 5 章，为了使读者掌握图像处理的概要，介绍数字图像处理的基本概念，二维线性系统和二维平面上的概率场，图像的数字化及图像的二维线性模型。图像的数学建模是图像的恢复与预测处理中非常重要的环节。

第 6 章，介绍数字图像的增强和恢复中的各种方法，并给出若干个示例。介绍增强图像的对比度、清晰度、平滑化和几何畸变校正等方法。详细介绍图像的边界提取和二值化方法。这些方法作为模式识别的预处理是非常重要的。介绍利用二维有限长单位冲激响应(FIR)滤波器或二维维纳(Wiener)滤波器恢复模糊图像或被噪声干扰图像的方法。因为从退化了的观测图像恢复原始图像需要对图像进行建模，所以讨论二维线性系统在本质上说是非常重要的。

第 7 章，分析数字图像的压缩问题，介绍变换编码方法和预测编码方法。特别对在 JPEG 和 MPEG 国际标准中所采用的，被广泛使用的二维离散余弦变换进行详细的论述。

第 8 章，详细介绍 CT 等医学图像处理中所必须的基于投影技术的数字图像的重构方法。

在本书中所论述的语音信号处理以及图像处理中的重要基本概念，在本系列书中的《数字信号处理基础理论》一书中也作了详细的论述，必要时读者可参考该书来理解本书中的有关内容。

编著者　谷萩隆嗣

目　录

第1章 数字语音信号处理概要

CONTENTS

用语言来传递信息是人类区别于其他动物的最显著特征。本章将介绍人类是如何生成语音以及感知语音的，并介绍在计算机上描述的语音生成模型。语音生成模型对人类以及计算机仿真都极为重要。

1.1 基于语音的信息传递

包括人类在内的动物都会以自己的方式来传递信息(information)，即通常用声音、视觉、气味以及其他的方式来进行信息的传递。例如，雌性犀牛会以外激素(pheromone)这种特殊物质告诉在远处的雄性犀牛当前自己所处的位置。此外，还有如大象等动物用人类听不到的低频声音来进行信息的传递。

但是，人类的语言(language)与其他动物的信息传递方式有本质上的差异[1]。其他动物的信息传递仅限于用某种信号来表达特定的需求和欲望。以某种信号或姿态来表达信息的方法在人类获得语言之前就已经存在了。如果不使用语言，而是单单使用某种信号来进行信息传递，要表示一个新的想法或是将某个事件进行组合并表示出来都是不可能的。

姿态是最不发达的传递信息的方法。用姿态可以表示“走”这一状态。可是要表示“今天将家里的狗赶走了”，“我打算明天走”，“想像走时候的情景”等使用“走”这个词的地方很多，但是这些信息如果不用语言都无法表达。即使对“中午我吃的是咖喱饭”这种简单的事件，如果不用语言也是无法表达的。

语言之所以可以传递复杂的信息，是因为用单词进行组合可以表达一个句子(syntax)。由于这种造句功能的产生，使人类可以无限地构造出各种语句来表达自己的思想。

此外，如“没发生什么事情”，“没有起来”这种具有否定含义的信

息的表示，只有语言才能够传递。

语言是将瞬间考虑到的事情用一个单词序列表达出来。姿态是将瞬间考虑的事情原封不动地形象地表达出来。比如，姿态无法准确地传递“5 时 46 分 51 秒”这个事件；要想表达这个事件必须使用语言。

人类在进化的过程中获得了说话的能力，给不同的发音(sound)赋以不同的语义(meaning)。但是仔细考虑一下，发音与语义之间不存在本质上的差异。下面我们来讨论一个问题：虽然耳朵和大脑连接的神经纤维只有两万根左右，为什么自然却选择了语音作为表达语言的手段。

用说话来表达语言的一大益处就是快速。从人类的历史来看，传递信息速度的快慢肯定是保住性命的一个重要的因素。说话的人一分钟可以表达由 200 到 300 个单词组成的简单语言，听的人一分钟则可以理解 500 个单词所组成的语言。

此外，在黑暗中用手语和姿态无法沟通，手被占用的时候也无法使用手语。即使你认为耳朵不是很出色的感觉器官，在这种情况下也可以捕捉与声音相关的语言。只有人类传递信息的时候可以使用**语音**(speech)，并且可以掌握并使用语言，**语音语言**(spoken language)的使用标志着人类与其他动物的不同。

1.2 人类的语音生成和语音感知

人类在说话的时候会在口中进行复杂的动作。舌头上下前后急促地运动，另外在口腔深处，气息或者穿过鼻腔，或者因受到阻碍而上下振动。嘴唇的形状则需要根据不同的发音进行不同的调整。这样，凭借舌、唇、上颚和喉等各个部位的动作组合的变化，人类可以发出的音约有4 000种。

但是，人在刚出生的时候只会发出其中一小部分的声音。在人类历史上所发明出来的所有语言都是由数十个元音和辅音形成的。音

节由/b/、/d/、/g/等辅音以及/a/、/i/等元音构成。元音传递的是音质，用元音可以唱歌、喊叫、拖长音和将声音传到远方等。在适当的环境下进行训练，人类可以发出约50种元音。但是，目前世界上最常用的各种语言只有五个元音，世界上各种语言中大约20%使用的是相同的五个元音。

辅音将各个元音隔开而使其听起来更加清楚，发元音/a/的时候嘴唇打开，发辅音/b/的时候嘴唇闭上。这样，在说话的时候，嘴会一张一闭，嘴的张闭形成了所有语言的基础。辅音和元音相互交叉发音是人类所特有的。当然，如捷克语那样，可以用辅音的各种组合来形成一个单词，或是辅音和元音的交替比较少的单词也存在。例如，表示鸵鸟的pstros的前四个音就是连续的四个辅音。

婴儿在出生六个月左右就可以发出各种声音，他们无论在什么地方都会尝试着发出各种声音。大约在出生六个月之后，他们就几乎可以发出全部的语音了。这种有意识地发出各种语音的行为具有很重要的意义。在世界上所发明的所有语言中使用的各种发音，归根结底是因为人类这种有意识地发出各种声音的行为。

孩子学会说话的必要条件是可以流利地发出各个音节，也就是说可以流利地将辅音和元音组合在一起发音。婴儿虽然不会说话，但是世界上所有的语言中通用的主要音节他们都会。

发出辅音及元音，以及听到并理解这些音是人类的基本生理能力。如果没有发音能力，人类就没有语言。

前面的音对后面接续的音有影响的发声组合(co-articulation)是人类语言极其重要的特征。所以，人类可以快速地说话。例如，在说话的时候不是发/b/、/a/，而是发/ba/的音。

此外，例如英语中的cool(/ku:l/)和kill(/kil/)这两个单词，在辅音/k/的后面所跟的元音不同，辅音/k/的发音也有微妙的变化。人类在发辅音之前，因为已经知道后续的元音，所以就预先将所需要的口形准备好了。

我们观察说话的整个过程，要控制发声器官的复杂动作需要大脑。黑猩猩仅仅是可以获得所有的可以发出的声音，但是由于它的大脑不具备控制发声器官的能力，因此就无法形成说话能力。

研究发现，人类的大脑中具有专门控制语言的区域。人类的右脑和左脑分别起着不同的作用，右脑主要具有视觉的感知以及进行空间处理的功能；左脑专门支配语言活动。

大脑所控制的发声器官首先是喉(larynx)，又称为喉节，在喝水的时候会上下运动。喉部有声带，声带是由肌肉和韧带组成，具有复杂的结构。其机能不仅仅是像单簧管的簧片那样振动，还可以在空气通过时非常敏捷地开关。也就是说，声带将由肺部出来的听不见的气流转换成声音发出。其他动物如果有声带，也可以发声并传递信息。

人类进化的最重要的部分之一就是喉节，即喉节可以下降到现在的位置上。如果喉节的位置如原始的尼安得特(Neanderthal)人那样在鼻腔的附近，就无法发出/i/、/u/、/ku/、/gu/等音。/i/和/u/是只有进化后的人类才可以发出来的音。

虽然人类可以发出数千种语音，但只是选择了其中的几百种形成了现在世界各地的语言。在世界的任何地方，都不会使用超过 100 种的语音。一般情况下，一种语言所使用的音只有 40 种左右。世界上的语言都是从数百个相同语音组中选出最普通的音来使用。即使某种语言中含有听不惯的语音，从表面来看也没有什么区别[1]。

1.3 语音生成模型

语音是由发出声响的能量源**声音激励源(声源)**(sound source)以及将声源的语音变形后赋予其声韵的**调音**(articulation)这两个要素构成。如图 1.1 所示，语音的形成可以分为声源和调音两大模式[2,3]。

我们来分析人类的发声器官(vocal apparatus)的特征，发声模型分为声道(vocal tract)模式与放射模式(radiation)，但是在进行**语音合**

图 1.1 声源和调音分开的语音生成模型

成(speech synthesis)和**语音识别**(speech recognition)时,则不一定要采用这种描述方式。

声道保持了在通过由喉、舌、上颚和鼻腔构成的声源通道上发声器官的复杂形状,并且其形状随时间变化,是具有特别复杂的声学特性(acoustic characteristics)的声音管道。从性能上讲,是线性(linear)、时变(time-varying)的共振及逆共振系统(resonant and antiresonant system)[4,5]。由口或唇发出声波的放射特性具有微分特性。

如果对声源进行粗略的划分,可以分成在给定时间区间内近似周期性的脉冲信号,以及在给定时间区间内为不规则信号两种情况,而这两种情况下的信号都具有很宽的频谱(spectrum),其**频谱包络**(spectral envelope)是非平稳的。但是如果将声源与调音的频谱包络综合起来考虑,认为声源具有平稳的频谱包络,从而可以建立**语音生成模式**(model for speech production)。

如果语音的生成系统是线性的,那么因为是时变系统,处理起来很不容易。但是,语音信号在短的时间区间内通常可以看成是定常的,所以可以把语音信号的当前值 $x(n)$ 看成是其过去值 $x(n-k)(k=1,2,\cdots,M)$ 和声源信号的当前值与过去值 $u(n-k)(k=0,1,2,\cdots,N)$ 的线性组合,即可表示为

$$x(n)=\sum_{k=0}^{N}b_k u(n-k)-\sum_{k=1}^{M}a_k x(n-k) \tag{1.1}$$

如果将 $x(n)$ 和 $u(n)$ 进行 z 变换之后分别表示为 $X(z)$ 和 $U(z)$,则有

$$X(z)=\sum_{k=0}^{N}b_k z^{-k}U(z)-\sum_{k=1}^{M}a_k z^{-k}X(z) \tag{1.2}$$

上式可以表示为