

国外信息科学经典教材

[语 音 合 成]

Progress in Speech Synthesis

Jan P. H. van Santen

(美) Richard W. Sproat 编

Joseph P. Olive

Julia Hirschberg

蔡莲红 杨鸿武 吴志勇 等译



机械工业出版社
CHINA MACHINE PRESS



附光盘

国外信息科学经典教材

语音合成

Jan P. H. van Santen

(美) **Richard W. Sproat** 编
Joseph P. Olive
Julia Hirschberg

蔡莲红 杨鸿武 吴志勇 等译



机械工业出版社

本书介绍了语音合成技术近年来取得的进展。全书分 8 部分共 46 章，主要内容包括语音信号处理和声源建模、语言学分析、发音器官合成与可视语音、拼接合成与自动切分、自然语音的韵律分析、韵律合成、评价与感知，并简单介绍了两个系统及其应用。

本书可供计算机工程、智能信息处理、电子工程等专业的研究人员或工程技术人员参考使用，也可作为相关专业研究生和本科生的教材。

Translation from the English language edition:

Progress in Speech Synthesis edited by Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg

Copyright © 1997 Springer-Verlag New York, Inc.

Springer-Verlag is a company in the Bertelsmann Springer publishing group

All Rights Reserved.

本书中文简体字版由施普林格出版公司授权机械工业出版社独家出版。版权所有，侵权必究。

本书版权登记号：图字 01-2003-3155

图书在版编目 (CIP) 数据

语音合成/ (美) 简·凡桑塔 (Jan P. H. van Santen) 等编; 蔡莲红等译. —北京: 机械工业出版社, 2005.3

书名原文: *Progress in Speech Synthesis*

国外信息科学经典教材

ISBN 7-111-15529-7

I. 语… II. ①简…②蔡… III. 语音合成—研究—进展 IV. H017

中国版本图书馆 CIP 数据核字 (2004) 第 111902 号

机械工业出版社 (北京市百万庄大街 22 号 邮政编码 100037)

责任编辑: 李利健 版式设计: 张世琴 责任校对: 陈延翔

封面设计: 刘吉维 责任印制: 石 冉

三河市宏达印刷有限公司印刷·新华书店北京发行所发行

2005 年 3 月第 1 版·第 1 次印刷

787mm×1092mm 1/16·26.5 印张·657 千字

0001—3000 册

定价: 55.00 元 (含 1CD)

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

本社购书热线电话 (010) 68326294

封面无防伪标均为盗版

译 者 序

语言处理是一个古老而又崭新的课题，它在不断地进步。从 1939 年纽约世界博览会展出的早期“声码器”，到 Haskins 实验室从 Frank Cooper 制造模式回放机开始所作的许多贡献，到 Fant 的名著，再到 Klatt 所作的巨大贡献。从这些研究以及其他一些重要贡献来看，语音合成与语音生成研究是密切相关的。如果我们想让一个机器像人一样讲话，我们最好先了解人类语音产生的机理。同样，为了解人类是如何产生语音的，我们可以这样测试我们的模型：让模型合成语音（或语音的某些特性），再将模型的输出与我们在自然语音中的测量结果进行比较。这就是 Mary Beckman 呈现给我们的思想。

本书是在第二届 ESCA/IEEE/AAAI 文语转换研讨会的基础上，由技术委员会成员来遴选候选文章，并邀请了一致认可的文章的作者来编写的。虽然本书介绍的是几年前的技术，但它涉及到的理论、方法、技术仍是需要我们认真学习和效仿的。

近些年，语音合成的研究在国内也成了一个关注的热点，其成果已走出实验室，在信息领域获得了应用。语音应用的需求也促进了技术的发展。然而相应的书籍还较少。

在 1997 年初，Bell 实验室的朋友给我们带来了这本书的英文版。它一直是清华大学计算机系人机语音交互组研究生的必读资料。不敢说我们都读懂了，理解了，但本书确实给我们上了一课。书中如此丰富的内容，如此多等待“进展”的课题，激发了研究生们的兴趣。

本书的翻译出版又给我们一次深入学习的机会，我们很高兴承担此任务。中译本的出版可以让国内更多的同行和学子了解语音合成，通过此书，让我们结交更多的朋友。

在本书的翻译出版过程中，清华大学计算机系人机语音交互组的研究生杨鸿武、吴志勇、蒋丹宁、倪昕、王志明、蔡锐、崔丹丹、郑敏、马磊参加了翻译工作，蔡莲红、杨鸿武、吴志勇审校了全书。由于译者的水平所限，书中难免出现不够准确或错误的地方，敬请读者指出。

译 者

出版说明

在人类迈入信息时代的今天，信息技术的应用无处不在，我国对信息技术的重视和鼓励也达到了空前的程度。信息技术的发展速度很快，可谓日新月异，尤其在一些发达的国家更是如此。我国信息技术起步较晚，但发展速度惊人，这正是改革开放的具体体现。在我国大力发展信息产业的今天，为了能融入国际的潮流和掌握最新的技术，从国外引进先进的知识和技术就显得格外重要。为此，我们决定引进一系列国外信息技术领域有代表性的优秀教材，将它们献给我们的学子、教师和 IT 业的有志之士，借此为我国的信息产业贡献一份微薄之力。

随着我国加入 WTO，国际间的竞争将越来越激烈，而国际间的竞争实际上就是人才的竞争、教育的竞争。为了加快培养具有国际竞争力的高水平技术人才，加快我国教育改革步伐，使我国的高等教育尽快与国际接轨，这就需要引进先进的教学思想和教学方法，而引进国外优秀的教材无疑是一种很好的途径。同时，引进国外的优秀教材也有利于提高我国自编教材的水平，让我们的教育工作者从中得到启发。

我们这套丛书遵循“新、优、特”的原则，做到知识新、质量优和内容有特点。这套丛书涵盖了计算机、通信、电子技术等领域，每一本书都是精心挑选，在某个领域或学科内具有很强的代表性和很高的价值，很多在国外也被作为大学的教科书，由国际知名的出版公司出版。在引进过程中，我们邀请有关专家对书稿的整体水平进行了评定；在翻译过程中，我们聘请国内相关领域内的有很高学术水平的专家和学者，以保证书籍的水平和质量，做到对读者负责，为读者着想。

相信这套丛书的出版对正在苦读和即将面临挑战的学子们会有很大的帮助和提高，也能让我们的教育工作者从中得到启发，同时对从事 IT 行业的工程技术和研究人员而言也是很难得的工具书。

机械工业出版社

前言

从文本到语音的转换、合成，包含了从输入文本到语音信号的各种计算。要满足这些计算需求，文语转换系统必须具备从对话结构的抽象语言学分析到语音编码的众多功能组件。

这一事实隐含了以下几个含义：首先，文语转换具有跨越多学科的性质。这一点可以从本书众多作者的背景得到证明，他们分别来自工程领域、语言学的多个领域、计算机科学、计算心理学和声学领域；第二，由于面对问题的复杂度差异很大，这些领域的研究进展是不平衡的。比如，为复杂的、多段的输入文本赋予无缝衔接的音高重音极为困难，而在其他音素都计算无误的情况下，生成音段时长并非十分困难；第三，总结 TTS 所有相关领域的研究惟一可行的办法是写一本多著者的书，因为任何个人，甚至研究小组都不可能具有足够广博的知识。

当然，这本书最重要的目标是邀请各相关领域的关键专家对各个领域作全面的评述和回顾。这将给读者展现一幅完整描绘 TTS 技术的图画：这一领域中所面临的挑战和解决思路是什么，研究者们正在向哪些方向前进。

本书的第二个重要目标是使读者能依据这些工作对最终结果做出判断——科技尖端的发展是令人鼓舞的，但是它是否能够真正产生高质量的合成语音？本书试图用两个方式来回答这个问题。第一，只要可能，我们都要求作者在他们所写的章节中收入主观评测结果。此外，书中包含有专门论述感知和评测问题的章节。第二，本书附带了一张包含有几个合成器合成样例的光盘。其中的成功与失败样例都十分有趣——尤其是后者，因为包含能够显示系统主要缺陷的样例通常是不太可能的。我们感谢 Christian Benoit 提出的这个建议。

在此，简短介绍一下本书的历史。1990 年，第一届 ESCA 文语转换研讨会在法国的 Autrans 召开。这次研讨会的组织者 Gerard Bailly 和 Christian Benoit 感到出版一本摘录长篇会议论文的书籍十分必要，因此《说话的机器》一书得以出版。在 1994 年，本书的编者组织了第二届 ESCA/IEEE/AAAI 文语转换研讨会，同样，我们感到有必要出版一本书籍来进一步适时、完整地展示当前的工作。为了保证本书的每一章节都能达到可能的最高质量，我们邀请了此次研讨会的技术委员会成员来遴选候选文章。本书编者也加入了他们的选择行列。而后我们邀请了那些得到一致认可的文章的作者来完成此书。

作为本书的编者，我们要感谢那些对本书的出版做出贡献的人们：技术委员会中参与遴选文章的成员、14 名匿名审稿人、Bernd Moebius 的精彩图示、Alice Greenwood 在编辑上的帮助、Mike Tanenblatt 和 Juergen Schroeter 处理了音频和视频文件、David Yarowsky 制作了

索引、Thomas von Foerster 和 Kenneth Dreyhaupt 在 Springer - Verlag 出版公司的高效工作、Cathy Hopkins 在管理上的帮助，以及 Bell 实验室对这一工作的支持和鼓励。

Jan P. H. van Santen

Richard W. Sproat

Joseph P. Olive

Julia Hirschberg

Murray Hill, New Jersey

1995 年 10 月

目 录

出版说明

译者序

前言

第一部分 信号处理和声源建模

第1章 简介: TTS 中声门声源建模新方法	1
1.1 声门声源建模简介	1
1.2 替换单脉冲激励	2
1.3 本部分指南	2
1.4 小结	3
参考文献	3
第2章 声门音位变体的合成	4
2.1 引言	4
2.2 实验数据	5
2.3 合成实验	6
2.3.1 材料	6
2.3.2 模型	8
2.3.3 方法	8
2.4 各个源参数对声门化的贡献	12
2.5 讨论	14
2.6 小结	15
参考文献	16
第3章 带有激励源参数动态控制的语音合成	17
3.1 引言	17
3.2 激励源模型	17
3.2.1 周期性激励	17
3.2.2 非周期性激励	18
3.2.3 LF 模型	18
3.3 分析过程	19
3.4 分析结果	21

3.4.1 语音材料	21
3.4.2 元音	21
3.4.3 元音边界	23
3.5 小结	23
参考文献	23
第4章 合成中语音信号非周期成分的修改	25
4.1 引言	25
4.2 语音信号分解	26
4.3 非周期成分的分析 and 合成	29
4.4 评价	31
4.5 语音修改	32
4.5.1 时间缩放	33
4.5.2 频谱修改	33
4.5.3 非周期成分脉冲的修改	33
4.5.4 周期/非周期信号比例的修改	33
4.6 讨论和结论	34
参考文献	34
第5章 文语转换中利用正弦模型的语音合成	36
5.1 引言	36
5.2 正弦模型概述	37
5.3 正弦分析	38
5.4 正弦合成	38
5.5 一般模型的简化	38
5.5.1 谐波正弦模型	38
5.5.2 系统幅度和相位	39
5.5.3 激励幅度和相位	40
5.6 简化正弦模型的参数	41
5.7 基频和时长修改	41
5.7.1 系统贡献	42
5.7.2 激励贡献	42
5.8 分析和再合成实验	42
5.9 结论	44
参考文献	44

第二部分 语言学分析

第6章 简介: TTS 合成系统中的文本分析	46
参考文献	47
第7章 语言无关面向数据的字音转换	48

7.1 引言	48
7.2 系统设计	49
7.2.1 对准	49
7.2.2 IG 树: 压缩和分类构造	50
7.3 相关方法	53
7.4 性能评价	54
7.4.1 连接	54
7.4.2 基于知识的语言学方法	54
7.5 结论	55
参考文献	56
第 8 章 语音合成中的全韵律结构	57
8.1 引言	57
8.2 系统结构	58
8.2.1 分析部分	58
8.2.2 语音解释部分 1: 时间的解释	60
8.2.3 语音解释部分 2: 参量解释	61
8.2.4 参数的产生与合成	63
8.3 多音节词	63
8.3.1 双音节	64
8.3.2 元音削弱和语音韵律	65
8.4 连续语音	67
8.5 总结	68
参考文献	69
第 9 章 一种非音段音位结构的定时模型	71
9.1 引言	71
9.2 音节联接及其在 YorkTalk 中的语音解释	72
9.2.1 参数化解释的原理	72
9.2.2 结构化的音位表示	72
9.3 节律描述及建模	73
9.4 YorkTalk 与自然语音及其他合成系统的比较	76
9.5 结束语	77
参考文献	77
第 10 章 一个完整的意大利语文语转换系统的语言分析	80
10.1 引言	80
10.2 形态分析	81
10.2.1 问题的定义	81
10.2.2 有关词典	82
10.2.3 形态分析器	84
10.3 语音转换	85

10.3.1 问题的定义	85
10.3.2 自动重音分配	86
10.3.3 开元音和闭元音	86
10.3.4 浊辅音和清辅音	87
10.4 形态-语法分析	87
10.4.1 预分析器	87
10.4.2 形态-句法分析器	88
10.4.3 语法解析器	89
10.5 性能评价	90
10.6 结束语	91
参考文献	91
第 11 章 记叙文中重音的语篇结构限制	93
11.1 引言	93
11.2 记叙文研究	93
11.2.1 分析	94
11.2.2 结果	94
11.3 基于语篇的重音功能解释	95
11.3.1 注意状态的建模	96
11.3.2 记叙文的语篇分析	98
11.4 重音的语篇功能	98
11.4.1 局部焦点的处理: 代词	98
11.4.2 全局焦点的处理: 显式形式	100
11.5 讨论	101
11.5.1 旧信息和新信息	101
11.5.2 主题划分	101
11.5.3 相对重要性	102
11.5.4 小结	103
11.6 结束语	103
参考文献	103
第 12 章 文语转换中的同形异音字消歧	106
12.1 引言	106
12.2 已有的方法	106
12.3 算法	107
12.4 歧义类的决策列表	113
12.4.1 类模型: 创建	113
12.4.2 类模型: 使用	114
12.4.3 类模型: 结合先验概率	114
12.4.4 罗马数字	114
12.5 评价	115

12.6 讨论和结论	116
参考文献	117

第三部分 发音器官合成与可视语音

第 13 章 简介：语音合成中“讲话的头”	119
参考文献	121
第 14 章 简介：发音器官合成与可视语音	122
参考文献	124
第 15 章 语音模型与语音合成	126
15.1 主题和一些例子	126
15.2 十五年的语调合成	127
15.3 时间模型	135
15.4 结束语	139
参考文献	139
第 16 章 基于伪发音器官参数合成语音片段的框架	145
16.1 引言	145
16.2 控制参数和映射关系	146
16.3 利用 HL 参数合成的例子	147
16.4 合成规则	149
参考文献	151
第 17 章 基于生物机械学和病理生理学的语音建模	152
17.1 引言	152
17.2 发音器官合成	152
17.3 一个有限元舌头模型	153
17.3.1 为软组织建模	153
17.3.2 舌头模型研究概要	154
17.4 控制器	156
17.5 结论	159
参考文献	159
第 18 章 会说话人脸的分析——合成与可懂度	161
18.1 引言	161
18.2 参数模型	161
18.3 视频分析	162
18.4 实时分析—合成	163
18.5 模型的可懂度	164
18.5.1 刺激数据的准备	164
18.5.2 整体可懂度	165
18.5.3 辅音混淆度	166

18.5.4 元音混淆度	167
18.6 视频分析	168
参考文献	169
第 19 章 可视语音合成中的 3D 嘴唇与下腭模型	170
19.1 引言	170
19.2 2D 嘴唇模型	170
19.3 3D 嘴唇模型	172
19.4 嘴唇模型的动画	174
19.5 下腭模型	175
19.6 嘴唇和下腭模型的动画	176
19.7 嘴唇和下腭模型的评价	176
19.8 结论	177
参考文献	177

第四部分 拼接式语音合成与自动切分

第 20 章 简介: 拼接式语音合成	179
第 21 章 德语拼接式语音合成中的混合基元结构	181
21.1 引言	181
21.2 自然语音概述	182
21.2.1 实验材料	182
21.2.2 浊音清化现象	182
21.2.3 同化现象	183
21.2.4 音节边界位置	183
21.2.5 元音前后的辅音	184
21.2.6 结论	184
21.3 基元结构与拼接规则	185
21.3.1 拼接方法	185
21.3.2 基元结构	185
21.3.3 基元定义	186
21.3.4 拼接规则	186
21.4 感知评估	187
21.4.1 配对比较实验	188
21.4.2 音段可懂度测试	190
21.5 小结	191
参考文献	192
第 22 章 拼接式语音合成中的韵律及基元选取	194
22.1 引言	194
22.2 切分及韵律标注	195

22.3 语音数据库基元定义	196
22.3.1 语音类别及音段样本	196
22.3.2 确定语音基元	196
22.3.3 数据库裁减	198
22.4 基于韵律的语音基元选取	198
22.5 实验及评估分析	200
22.5.1 实验 1: 使用全部数据库进行测试	200
22.5.2 实验 2: 使用裁剪后的数据库进行测试	202
22.6 讨论	202
22.7 结论	203
参考文献	204
第 23 章 双音子的优化拼接	206
23.1 引言	206
23.2 未优化的双音子集合	206
23.3 不匹配度度量方法	207
23.3.1 简单帧不匹配度	207
23.3.2 考虑帧以及回归系数的不匹配度	209
23.3.3 基于帧窗口线性拟合的不匹配度	211
23.3.4 给定时长的最小不匹配度	212
23.4 实验评估	212
23.4.1 总体考虑	212
23.4.2 感知实验	213
23.5 结论	214
参考文献	214
第 24 章 应用于拼接基元选取的自动语音切分	216
24.1 引言	216
24.2 自动标注算法	216
24.2.1 音位结构学模型	216
24.2.2 时长模型	217
24.2.3 音素声学模型	217
24.2.4 切分算法	217
24.2.5 训练算法	218
24.3 切分实验	218
24.4 结论	220
参考文献	221
第 25 章 Aligner: 使用 Markov 模型进行文语对齐	222
25.1 引言	222
25.2 Aligner 的操作	224
25.2.1 产生语音序列	224

25.2.2 文语对齐	225
25.3 评估实验	225
25.4 讨论和结论	228
参考文献	229

第五部分 自然语音的韵律分析

第 26 章 简介：韵律分析：一条双重途径？	230
第 27 章 简介：自然语音的韵律分析	232
第 28 章 利用统计分析自动提取 F_0 控制规则	234
28.1 引言	234
28.2 自动提取 F_0 控制规则的算法	234
28.2.1 规则提取过程概述	234
28.2.2 F_0 包络分解	235
28.2.3 统计规则提取	237
28.3 F_0 控制规则提取实验	237
28.3.1 语音数据及参数提取条件	237
28.3.2 用以建模的语言学参数	238
28.3.3 F_0 控制规则解释	239
28.4 小结	241
参考文献	243
第 29 章 语音合成中音高包络规格化方法的比较研究	245
29.1 引言	245
29.2 基于音调感知的自动规格化方法	246
29.2.1 理论基础	246
29.2.2 音调感知和韵律分析	247
29.2.3 算法描述	248
29.2.4 讨论	250
29.3 手工直线规格化方法	251
29.4 感知和直线规格化方法的比较	252
29.4.1 两种方法的差异	252
29.4.2 感知实验	253
29.5 结论	255
参考文献	255
第 30 章 z-score 模型中的停顿生成	257
30.1 引言	257
30.2 节奏和感知中心	259
30.2.1 关于每个音节一个参照点的争论	259
30.2.2 周期的吸引者	259

30.2.3 感知中心的声学互相关性	259
30.2.4 感知中心在感知调整上的重要性	260
30.3 Campbell 模型	260
30.4 Barbosa – Bailly 模型	260
30.4.1 IPCG 时长预测	260
30.4.2 语料	262
30.4.3 在重新分配算法中加入停顿现象	263
30.4.4 自动学习	264
30.5 感知测验	265
30.6 小结	266
参考文献	266
第 31 章 贝尔实验室里文语转换系统的时长研究	270
31.1 引言	270
31.2 数据库	270
31.3 时长模型	274
31.3.1 元音	275
31.3.2 摩擦音	276
31.3.3 爆发和送气	276
31.3.4 结尾部分	277
31.3.5 时长估计	277
31.4 讨论	278
31.4.1 补偿效应	278
31.4.2 句尾效应的缺失	279
31.5 小结	280
参考文献	281
第 32 章 德语语调曲线的合成	283
32.1 引言	283
32.2 语调模型	284
32.3 参数估计	286
32.4 基于规则的 F_0 合成	287
32.5 感知实验	288
32.6 小结	291
参考文献	292
第 33 章 说话风格对基频包络参数影响的研究	294
33.1 引言	294
33.2 语音素材	294
33.3 基频包络参数分析	295
33.3.1 提取基频包络参数	295
33.3.2 不同风格间的 F_0 参数比较	296

33.4 说话风格转换	298
33.4.1 转换到其他说话风格的转换规则	298
33.4.2 转换语音的评价实验	299
33.4.3 评价测试结果	299
33.5 小结	301
参考文献	302

第六部分 韵律合成

第 34 章 简介：文本与韵律	303
34.1 引言	303
34.2 控制 TTS 系统中的韵律	303
34.3 TTS 中语音风格的控制	304
34.4 抽象语音结构和语音事实	304
参考文献	305
第 35 章 简介：语调的语音表现	306
35.1 引言	306
35.2 使用自下而上的分析建立音位表现	306
35.3 使用自上而下的分析建立语音模型	307
35.4 语音表现和认知功能	307
35.5 韵律原型	307
35.6 结论	308
参考文献	309
第 36 章 瑞典语语调生成的词典语法信息计算提取	311
36.1 引言	311
36.2 瑞典语韵律结构	313
36.2.1 韵律词	313
36.2.2 韵律短语	314
36.2.3 韵律话语	315
36.2.4 末位延长与静音间隔	315
36.3 韵律结构组件的设计	316
36.4 性能	319
36.5 技术数据	319
36.6 小结	320
参考文献	320
第 37 章 用 TTS 符号输入进行的韵律变量参数控制	322
37.1 KIM - Kiel 声调模型	322
37.1.1 概述	322
37.1.2 重音	322