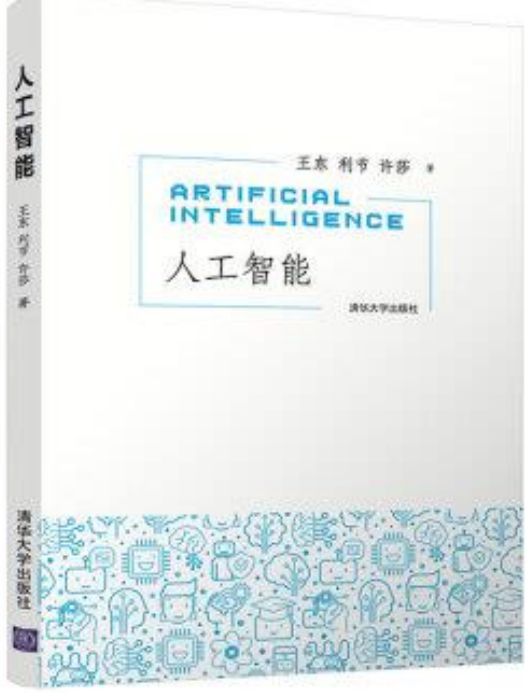
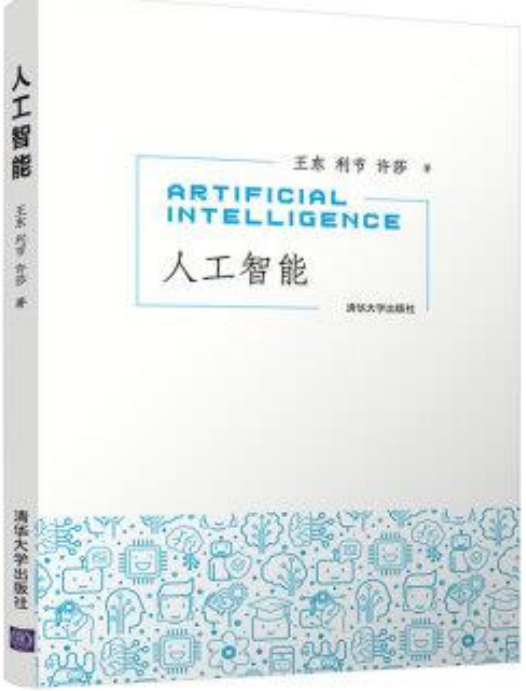
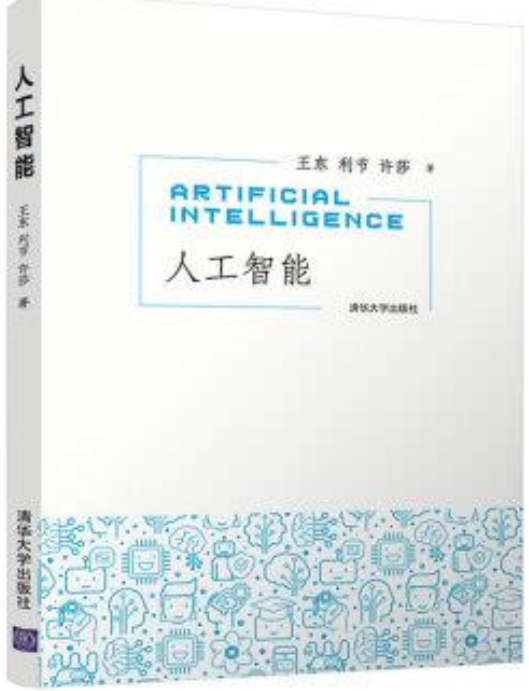


倾听你的声音

利节
改编自汤志远版本

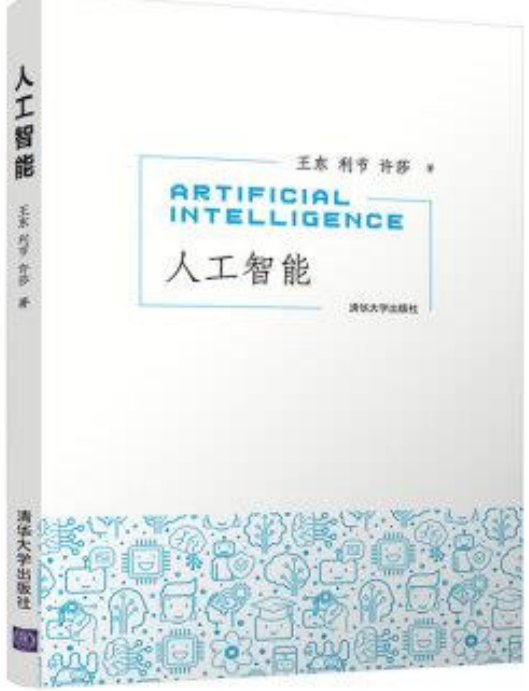

人工智能
王东 利节 许莎 著
ARTIFICIAL
INTELLIGENCE
人工智能
清华大学出版社

腾讯视频


人工智能
王东 利亨 许莎 著
ARTIFICIAL INTELLIGENCE
人工智能
清华大学出版社

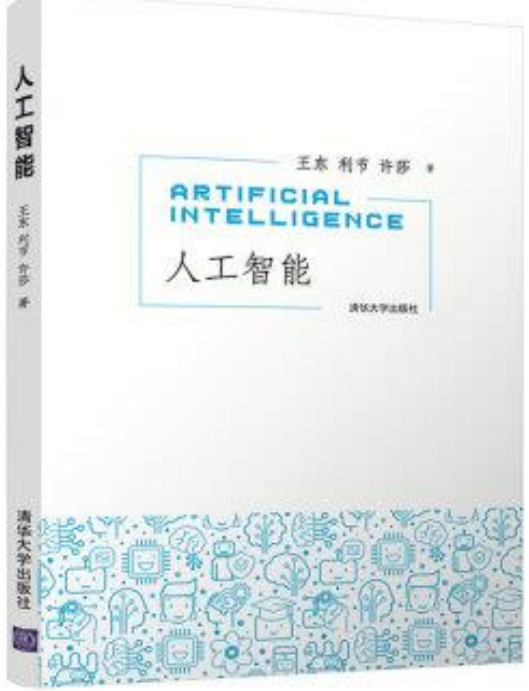
17:38
0.02K/s
16%
文件传输助手
2"
1"
1"
1"
1"
使用听筒模式
收藏
提醒
转换为文字（仅普通话）
删除
更多
2"
按住 说话

17:41
0.02K/s
15%
文件传输助手
1"
1"
1"
1"
2"
语音输入法
发送
倾听中...
上滑选择语种
符
123
空格
英/中



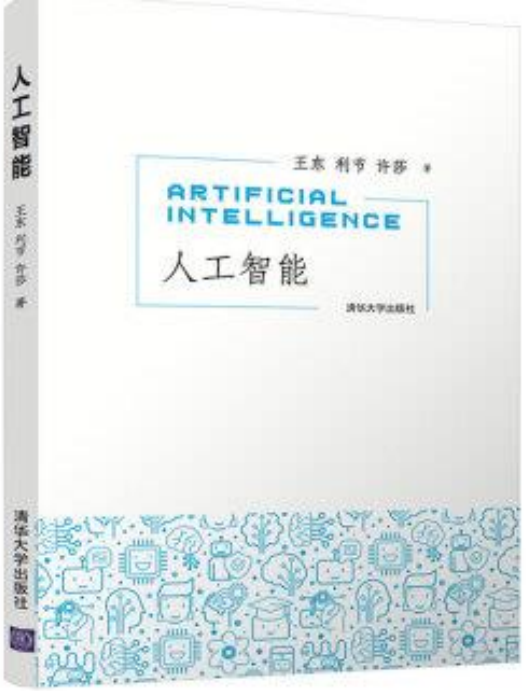
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用

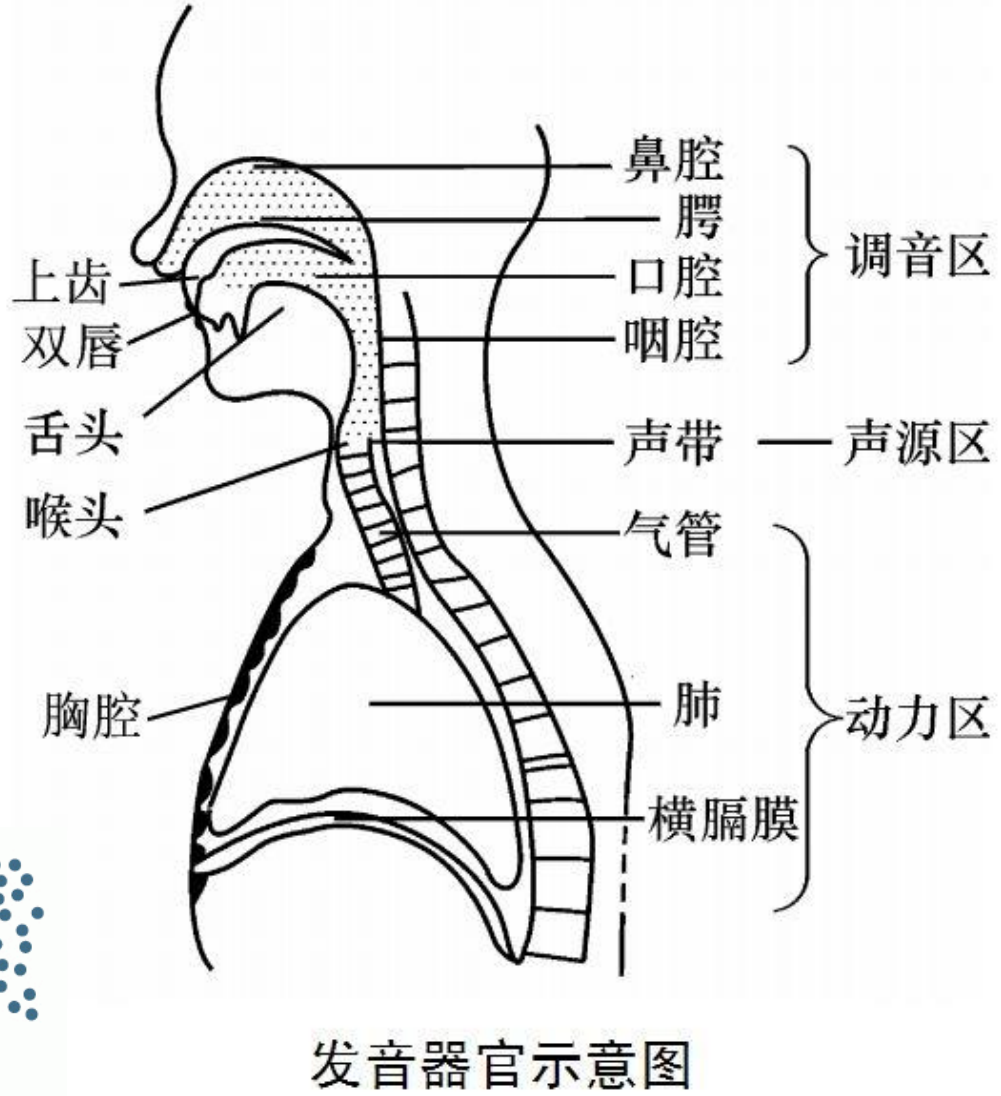


目 录

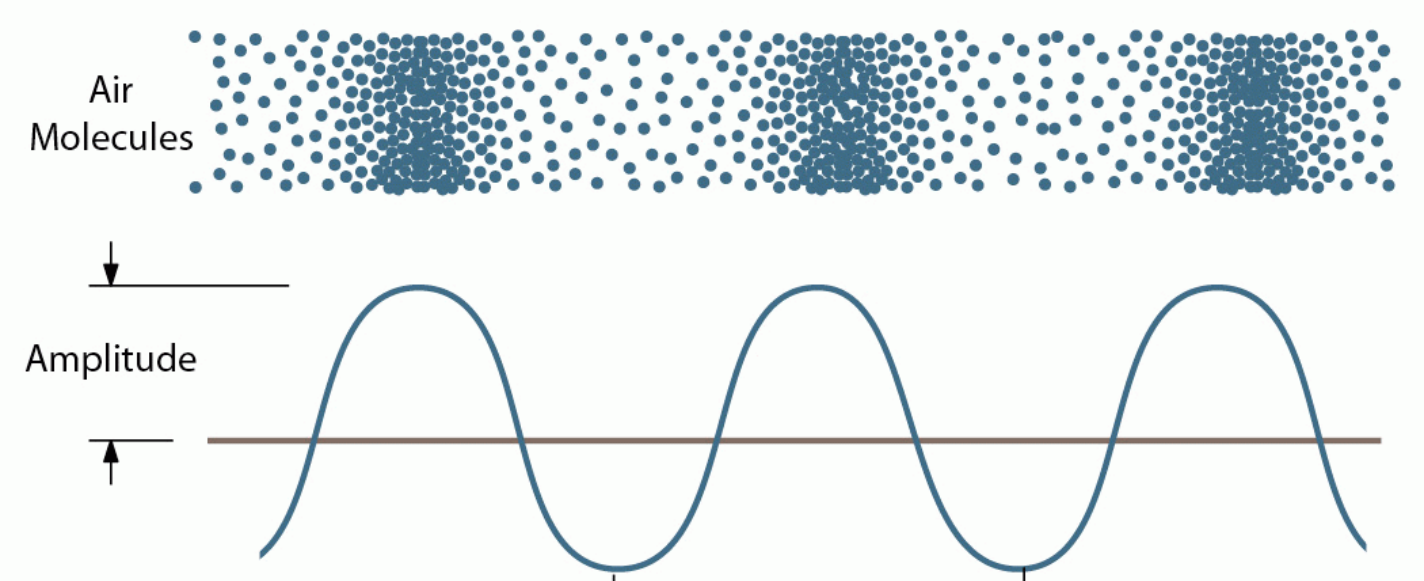
- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用

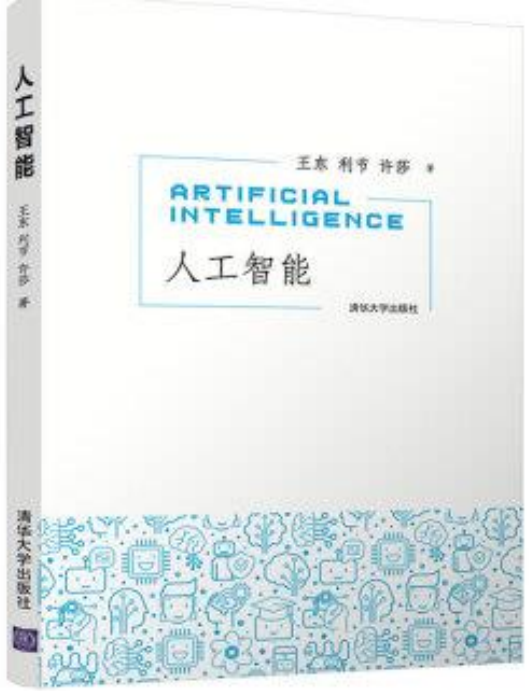


声音是什么



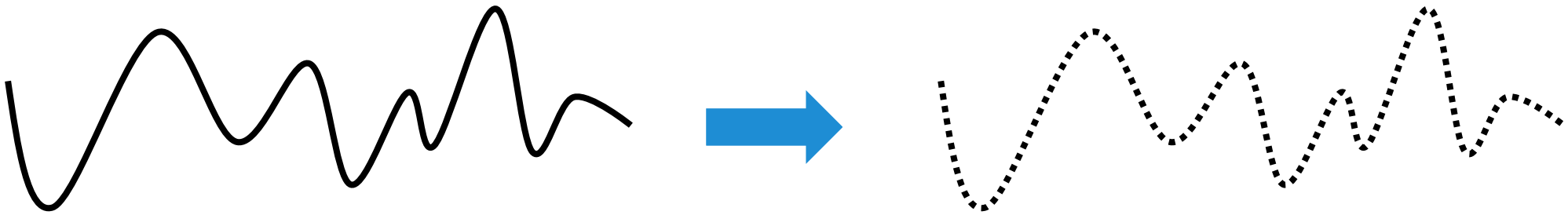
发音器官示意图





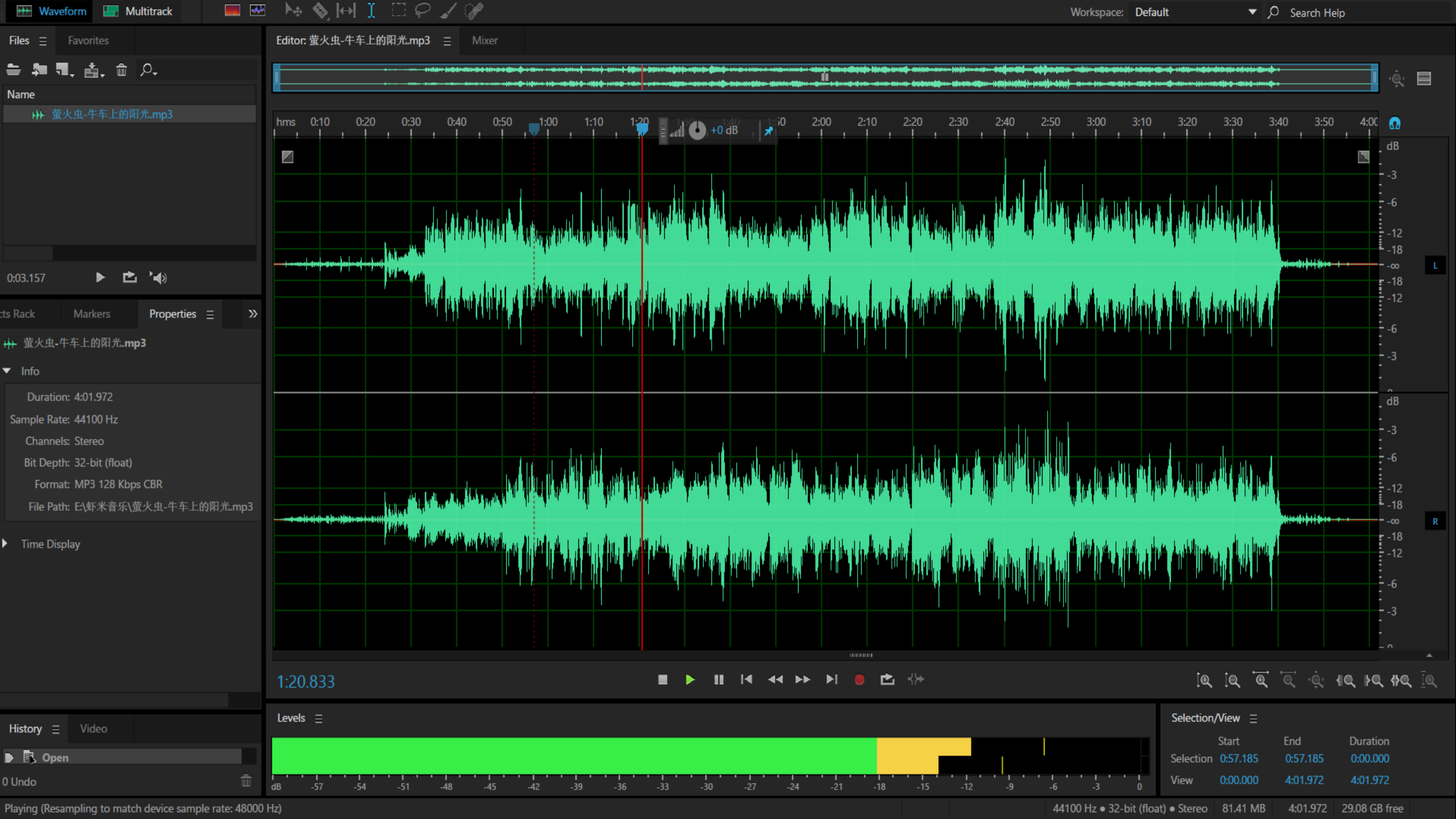
计算机眼中的语音

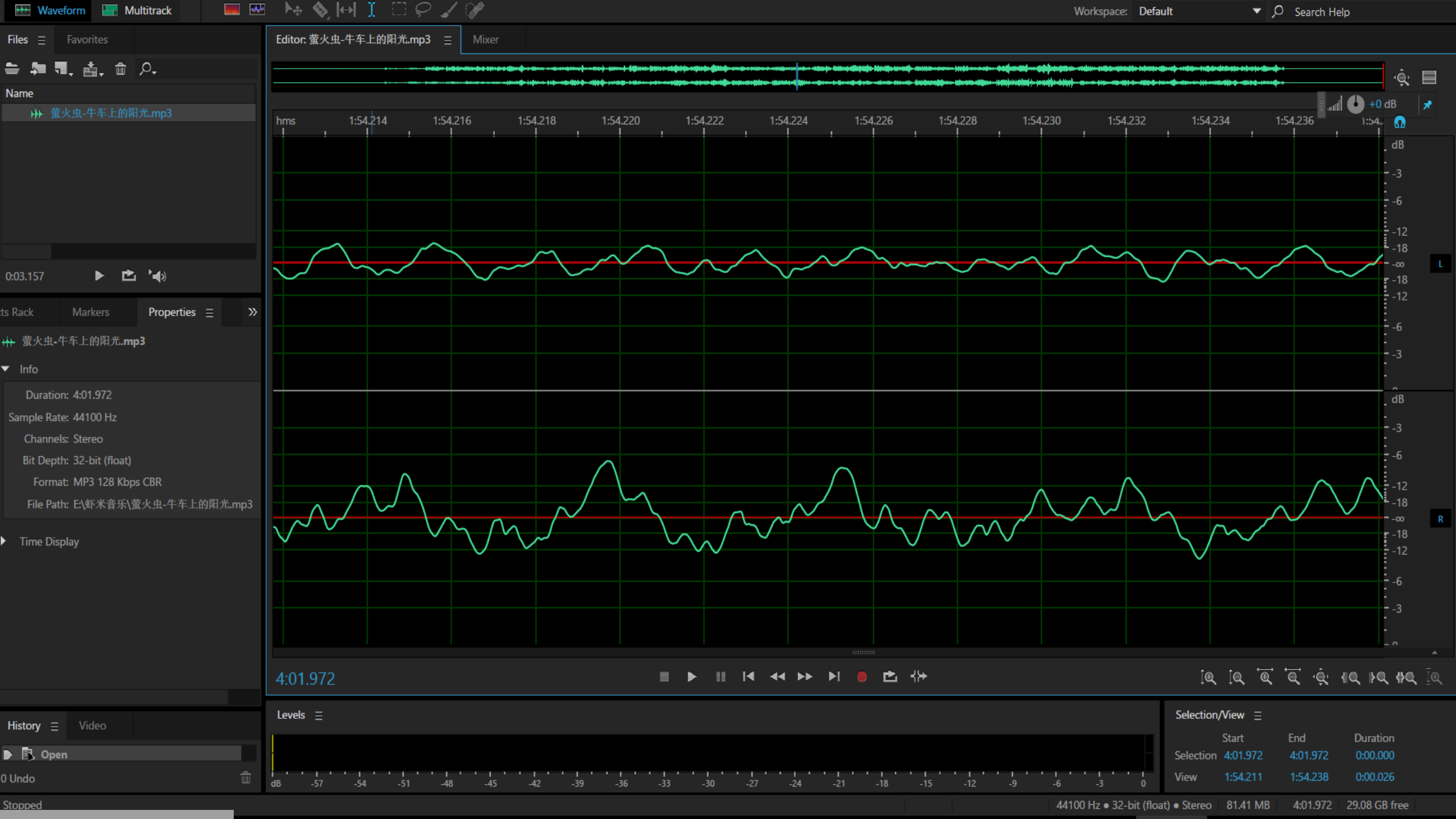
模数转换

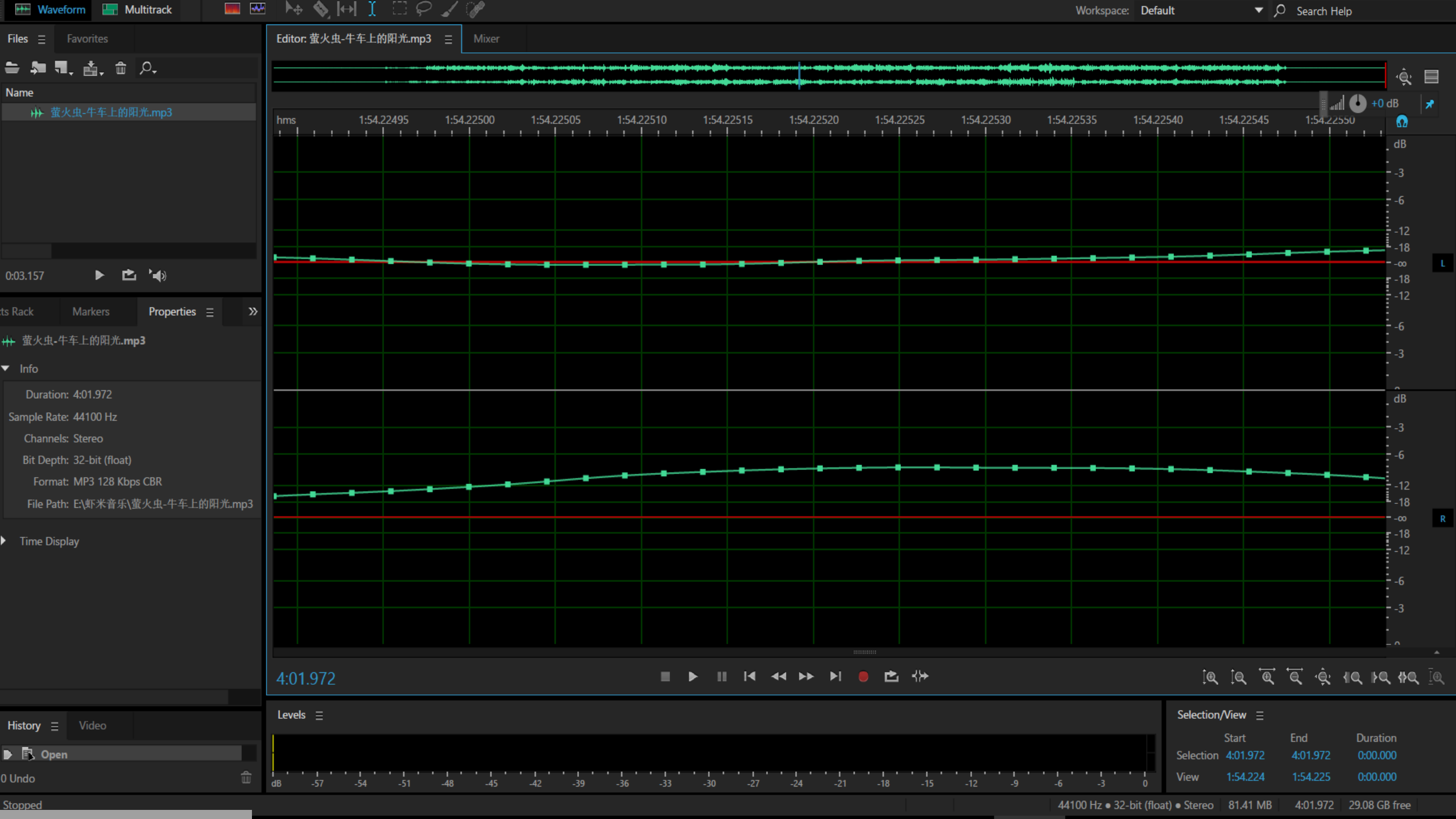


模拟信号

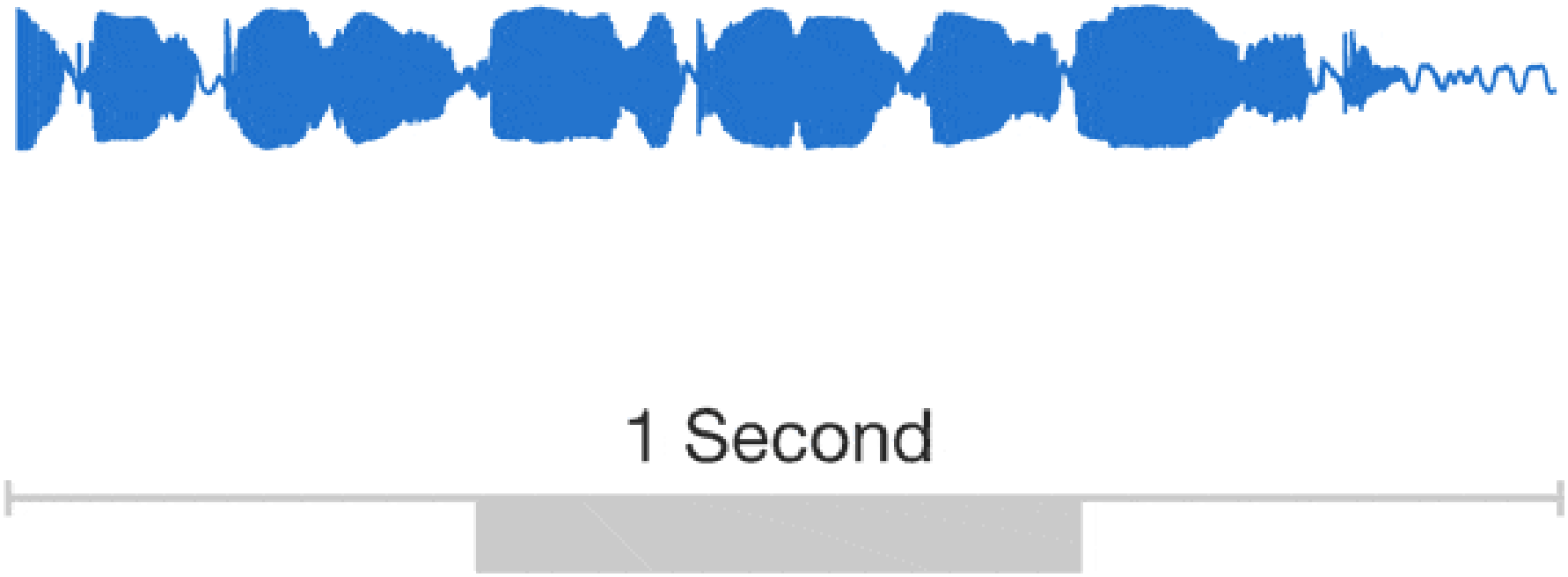
数字信号

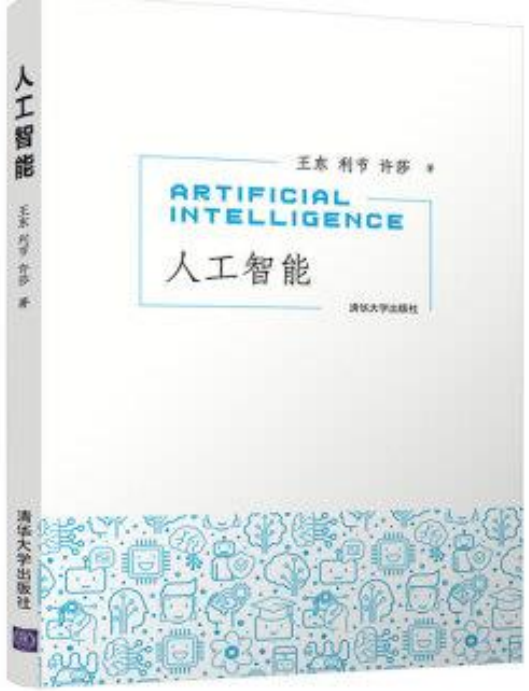

Waveform
Multitrack
Workspace: Default
Search Help
Files
Favorites
Name
萤火虫-牛车上的阳光.mp3
0:03.157
Editor: 萤火虫-牛车上的阳光.mp3
Mixer
hms 0:10 0:20 0:30 0:40 0:50 1:00 1:10 1:20 2:00 2:10 2:20 2:30 2:40 2:50 3:00 3:10 3:20 3:30 3:40 3:50 4:00
+0 dB
dB -3 -6 -12 -18 -∞ -18 -12 -6 -3
L
R
1:20.833
Markers
Properties
萤火虫-牛车上的阳光.mp3
Info
Duration: 4:01.972
Sample Rate: 44100 Hz
Channels: Stereo
Bit Depth: 32-bit (float)
Format: MP3 128 Kbps CBR
File Path: E:\虾米音乐\萤火虫-牛车上的阳光.mp3
Time Display
History
Video
Open
0 Undo
Levels
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
Selection/View
Start End Duration
Selection 0:57.185 0:57.185 0:00.000
View 0:00.000 4:01.972 4:01.972
Playing (Resampling to match device sample rate: 48000 Hz)
44100 Hz • 32-bit (float) • Stereo
81.41 MB
4:01.972
29.08 GB free


Waveform
Multitrack
Workspace: Default
Search Help
Files
Favorites
Name
萤火虫-牛车上的阳光.mp3
0:03.157
Editor: 萤火虫-牛车上的阳光.mp3
Mixer
+0 dB
hms
1:54.214
1:54.216
1:54.218
1:54.220
1:54.222
1:54.224
1:54.226
1:54.228
1:54.230
1:54.232
1:54.234
1:54.236
dB
-3
-6
-12
-18
-∞
L
R
Markers
Properties
萤火虫-牛车上的阳光.mp3
Info
Duration: 4:01.972
Sample Rate: 44100 Hz
Channels: Stereo
Bit Depth: 32-bit (float)
Format: MP3 128 Kbps CBR
File Path: E:\虾米音乐\萤火虫-牛车上的阳光.mp3
Time Display
4:01.972
History
Video
Open
0 Undo
Levels
dB
-57
-54
-51
-48
-45
-42
-39
-36
-33
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
Selection/View
Start
End
Duration
Selection 4:01.972 4:01.972 0:00.000
View 1:54.211 1:54.238 0:00.026
Stopped
44100 Hz • 32-bit (float) • Stereo
81.41 MB
4:01.972
29.08 GB free


Waveform
Multitrack
Workspace: Default
Search Help
Files
Favorites
Name
萤火虫-牛车上的阳光.mp3
0:03.157
Editor: 萤火虫-牛车上的阳光.mp3
Mixer
+0 dB
hms
1:54.22495
1:54.22500
1:54.22505
1:54.22510
1:54.22515
1:54.22520
1:54.22525
1:54.22530
1:54.22535
1:54.22540
1:54.22545
1:54.22550
dB
-3
-6
-12
-18
-∞
L
R
Markers
Properties
萤火虫-牛车上的阳光.mp3
Info
Duration: 4:01.972
Sample Rate: 44100 Hz
Channels: Stereo
Bit Depth: 32-bit (float)
Format: MP3 128 Kbps CBR
File Path: E:\虾米音乐\萤火虫-牛车上的阳光.mp3
Time Display
4:01.972
History
Video
Open
0 Undo
Levels
dB
-57
-54
-51
-48
-45
-42
-39
-36
-33
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
Selection/View
Start
End
Duration
Selection 4:01.972 4:01.972 0:00.000
View 1:54.224 1:54.225 0:00.000
Stopped
44100 Hz • 32-bit (float) • Stereo
81.41 MB
4:01.972
29.08 GB free

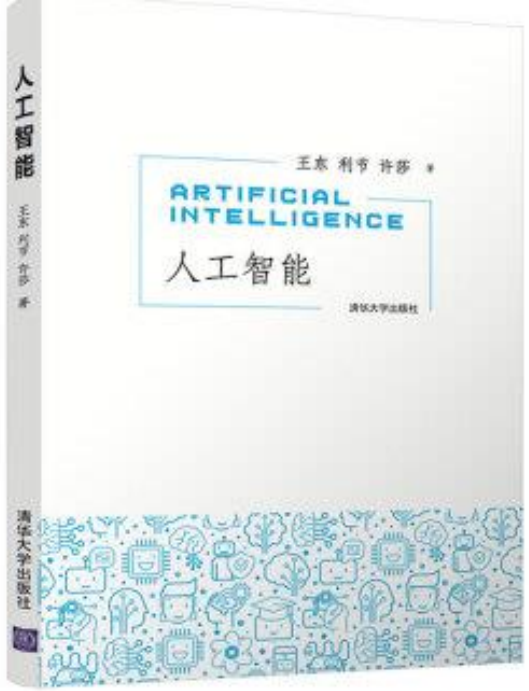
计算机眼中的语音





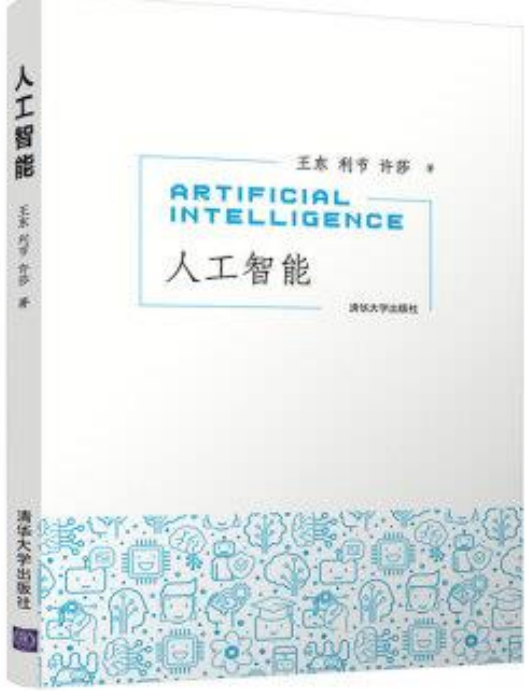
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用

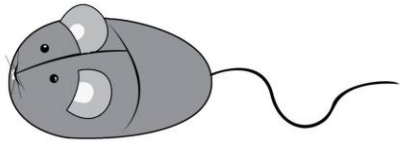


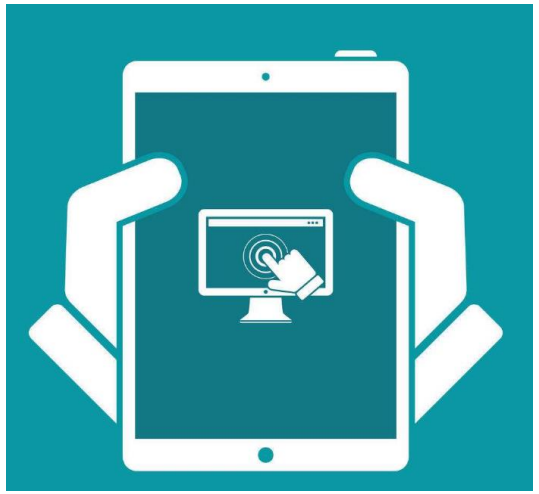
语音识别概述

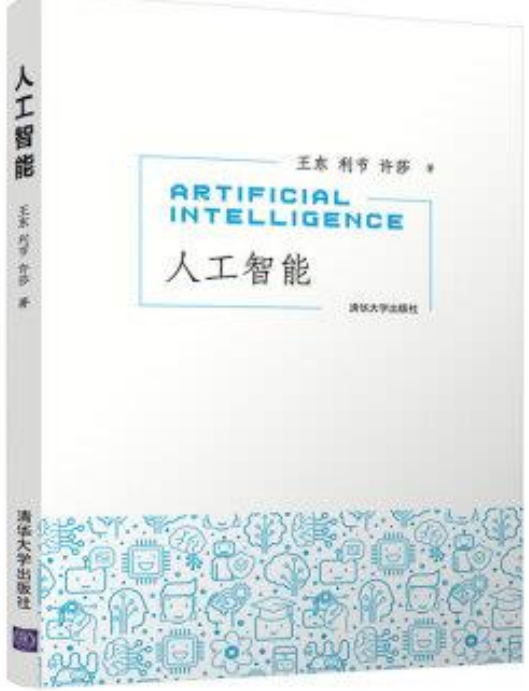
语音信号中包含丰富的信息，如果能将这些信息一一解析出来，将极大提高人类和机器交互的效率。从语音信号中解析各种信息的过程称为语音信息处理，包括语音识别（判断发音内容）、说话人识别（判断说话人身份）、情绪识别（判断发音人情绪）、语种识别（判断所用的语言）等。在众多语音信息处理任务中，研究最多、发展最成熟的是语音识别。

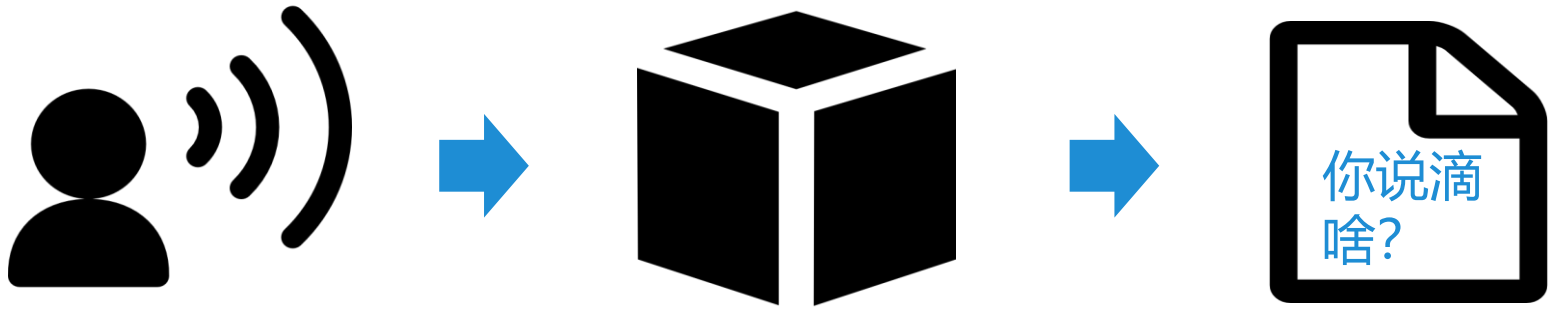


人机交互方式?

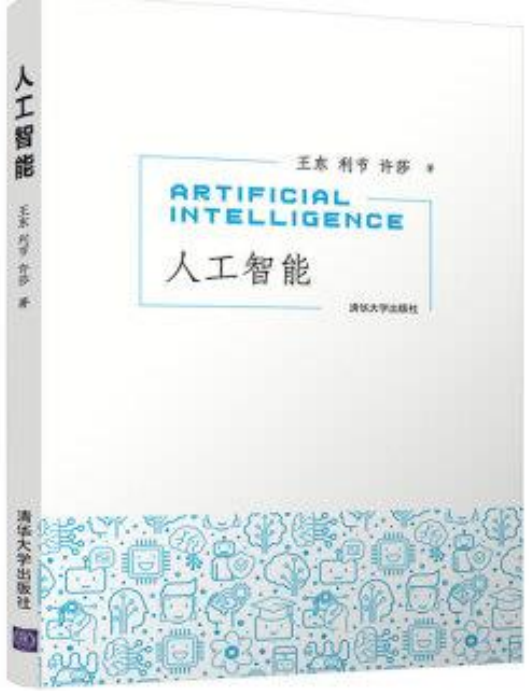








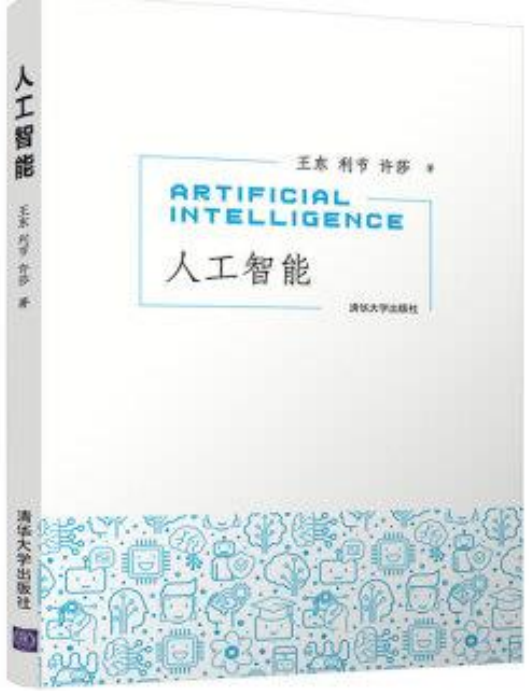
Automatic Speech Recognition (ASR)



Siri

Google ASSISTANT

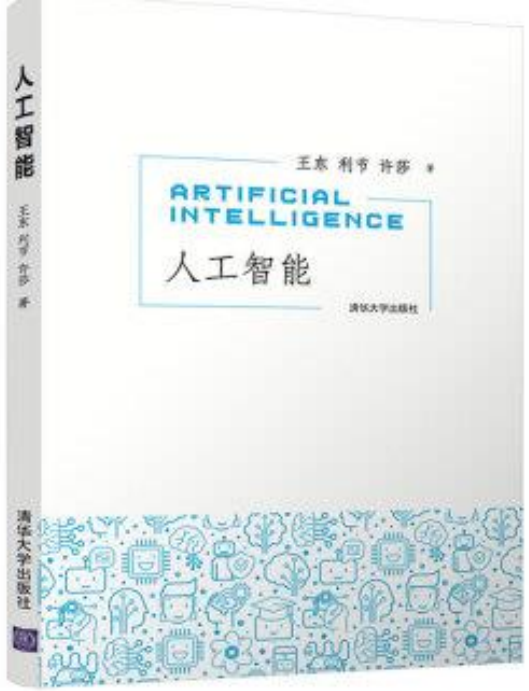
Cortana



语音识别概述

1)什么是语音识别

语音识别(Automatic Speech Recognition, ASR)是指利用计算机将语音转换成文字的过程，简单来说，就是让机器听懂我们在说什么。在实际应用中，语音识别经常与自然语言理解（理解句子的语义）、自然语言生成（生成文本句子）和语音合成（由文本生成声音）等技术结合在一起，形成一个基于语音的完整的人机交互系统。

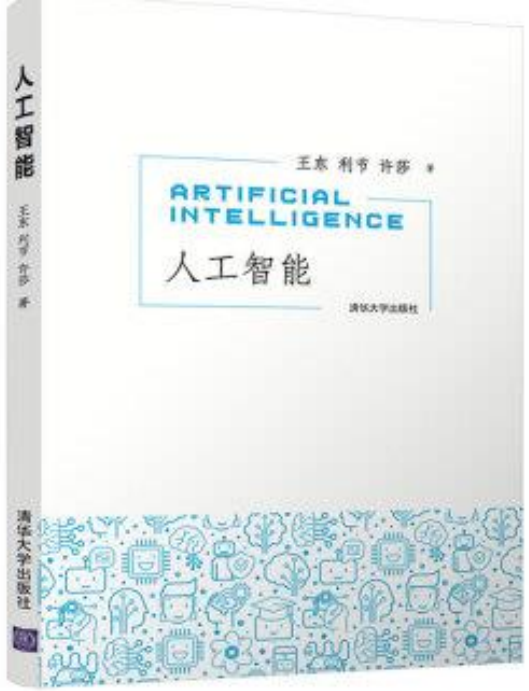


语音识别概述

2)语音识别简史

（a）知识积累时代（1930-1950）

这一时期计算机还没有出现，科学家们的主要研究内容是语音信号的短时分析，提出了短时频谱分析、滤波器组方法（Filter Bank）、线性预测分析等技术。这些技术为深入解析语音信号打下了基础。

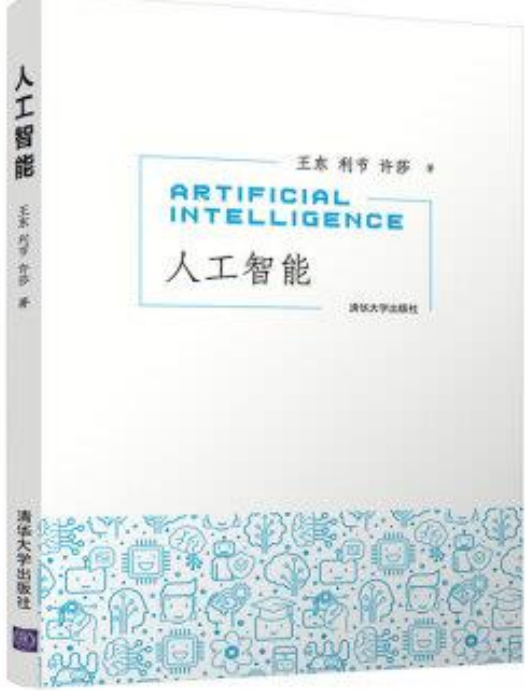


语音识别概述

2)语音识别简史

（b）模板匹配时代（1950-1980）

这一时期研究者可以利用简单的模板匹配算法识别少量单词和短语。所谓模板匹配，是指为每个单词或短语保留若干个发音样例作为模板；对一个待识别语音，将它与所有模板进行匹配，匹配度最大的模板所对应的单词或短语即可作为识别结果。



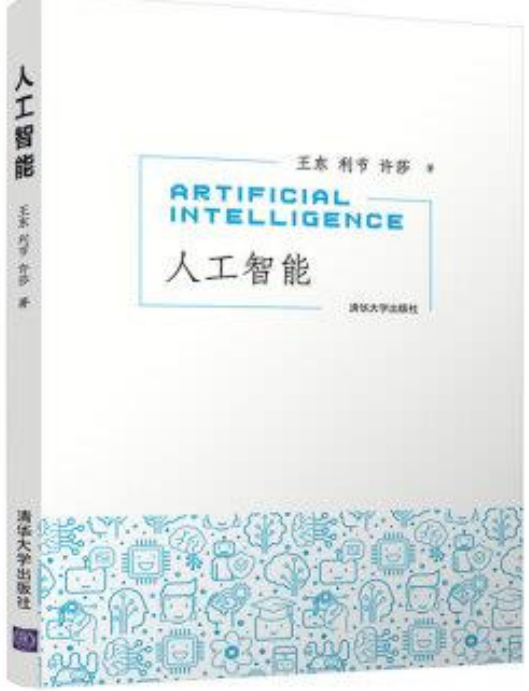
语音识别概述

2)语音识别简史

（c）统计模型时代（1980-2000）

基于DTW 的模板匹配方法有两个缺陷：首先，模板匹配仅可识别孤立的单词或短语，很难扩展到包含较多单词的长句子；其次，人类的发音具有变动性（如一个音素a 在每次发音时生成的语音信号不可能完全一致），这些变动性很难用有限的几个模板来描述。

最成功的统计模型是基于混合高斯的隐马尔可夫模型（GMM-HMM）。

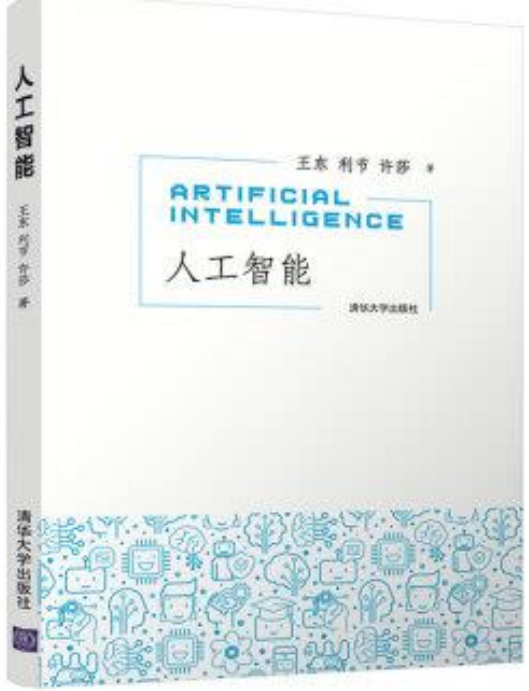


语音识别概述

2)语音识别简史

（d）机器学习时代（2000-2010）

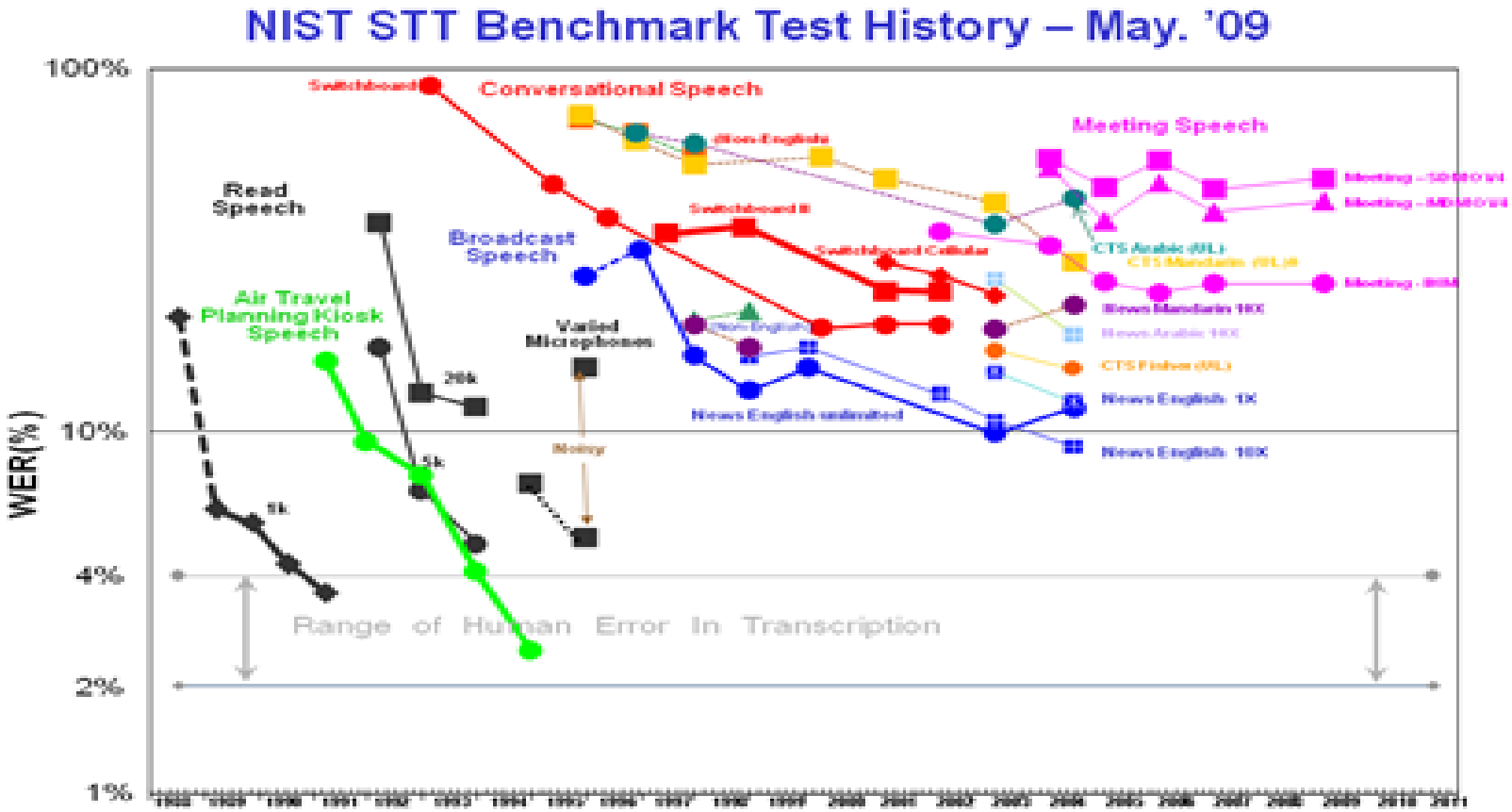
二十一世纪的前十年是语音识别技术的打磨期。这一时期没有划时代的技术革新，但出现了一些新的趋势，这些趋势为下一次技术革命打下了基础。首先，大量真实数据开始运用在模型训练中，极大提高了系统的实用性；另一方面，机器学习技术被广泛应用，以区分性训练为代表的新方法将GMM-HMM 系统的性能发挥到了极致。这一时期，语音识别开始走向实用化，出现了Nuance 这样的专业语音技术公司。

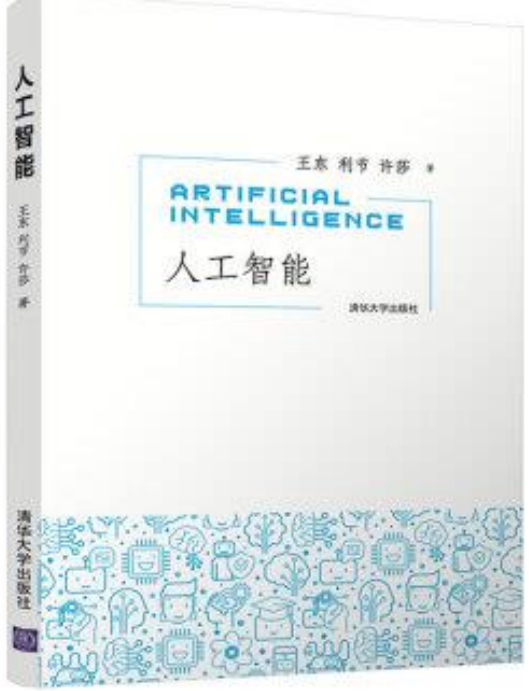


语音识别概述

2)语音识别简史

（d）机器学习时代（2000-2010）



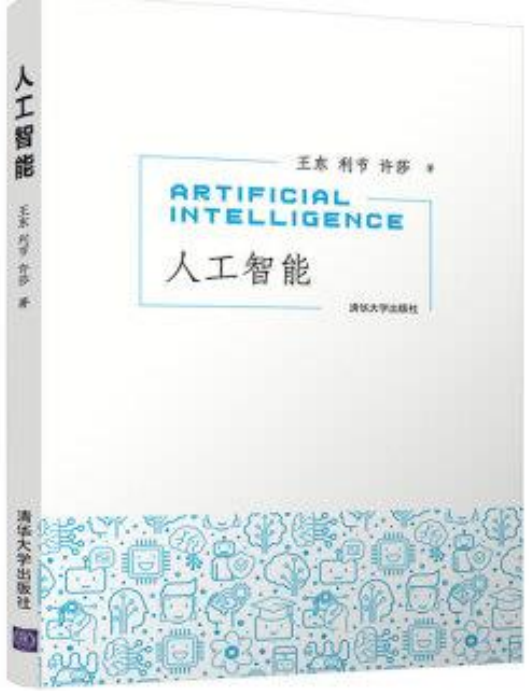


语音识别概述

2)语音识别简史

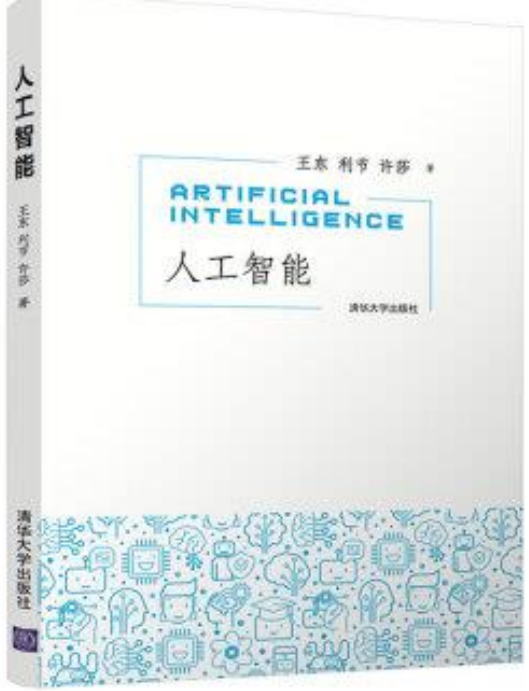
（e）深度学习时代（2011-）

变革发生在2009年。在这一年的NIPS Workshop上，Mohamed和他的导师Hinton 发表了一篇论文，报告了他们利用深度神经网络（DNN）进行声学建模的结果。Mohamed 发现，利用DNN，在TIMIT 数据集上可得到23% 的音素错误率，显著好于传统的GMM-HMM 模型。这一结果点燃了人们对深度神经网络的热情。在接下来的几年里，微软、IBM、谷歌等公司对DNN 模型进行了深入探索，尝试了各种模型结构和训练方法，取得了前所未有的性能提升。特别是在实际应用场景下，基于大规模数据训练的DNN 模型在某些领域甚至已经超过人类的识别水平。今天，DNN 已经成为语音识别的主流模型，统治业界30 年的GMM-HMM 模型已然退出历史舞台。



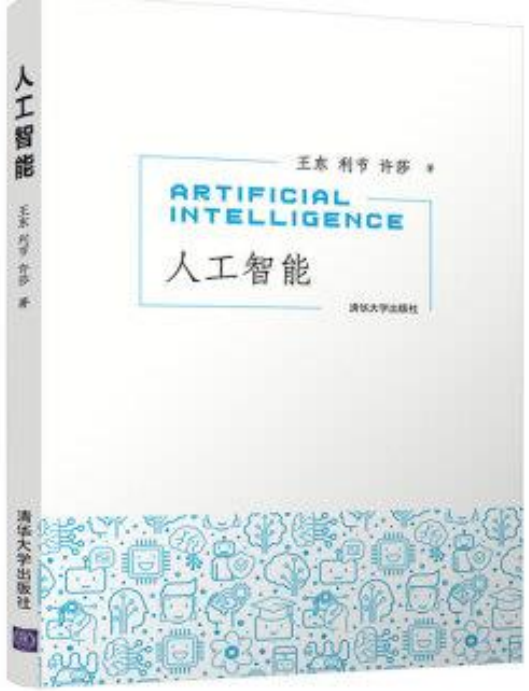
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用



基于GMM-HMM的语音应用

在统计模型时代，语音识别系统已经发展得非常完善，形成了一套完整的识别框架。和人脸识别系统一样，这一框架包括两个主要组成部分：特征提取和模式匹配。和人脸识别不同的是，语音识别的模式匹配是一个序列匹配问题，一般表现为一个搜索过程。通常将这种基于搜索的模式匹配过程称为解码，实现解码的模块称为解码器。在解码时，需要利用两个信息源：一个是描述每个音素如何发音的声学模型，另一个是描述单词组合规律的语言模型

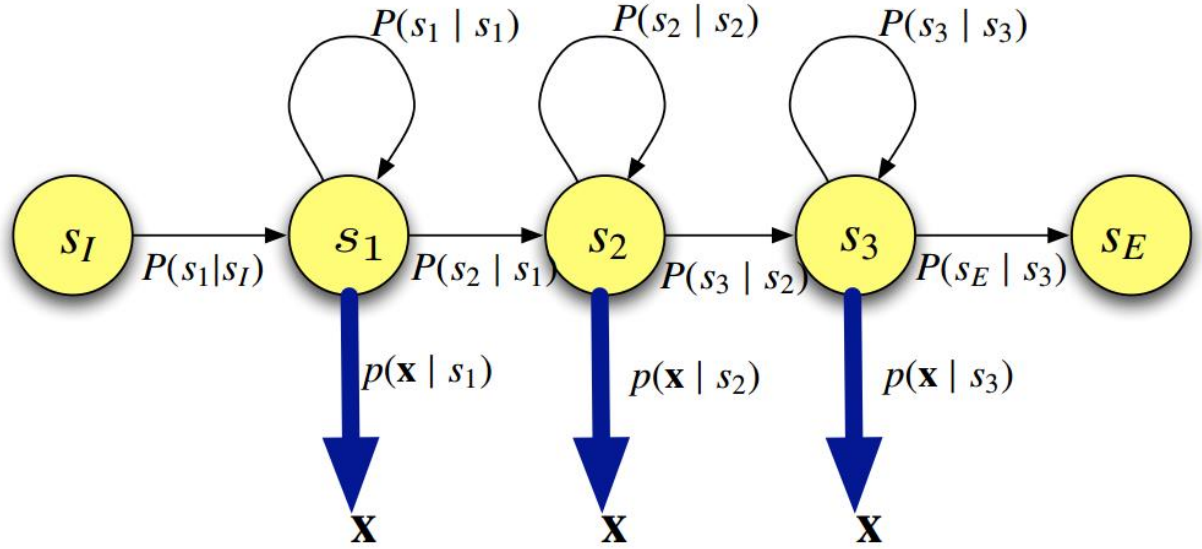


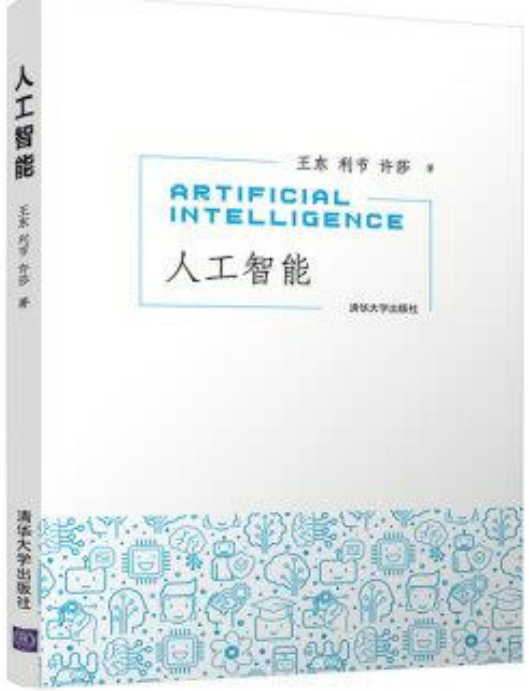
基于GMM-HMM的语音应用

1）HMM

隐马尔科夫模型（Hidden Markov Model，HMM），它用于处理时间序列数据，即样本之间有时间序列关系的数据。

比如语音识别，给你一段音频数据，需要识别出该音频数据对应的文字。这里音频数据就是观测变量，文字就是隐藏变量。



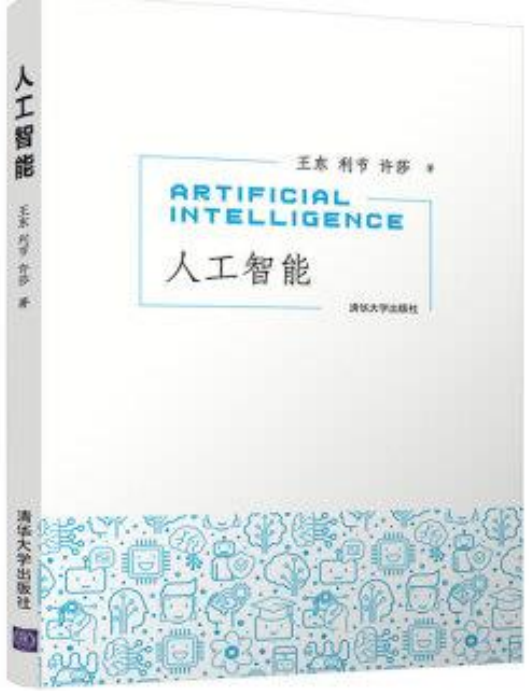


基于GMM-HMM的语音应用

1）HMM

对HMM而言，一般观测序列越长，推断越准。比如，我想输入“从一段字符序列推断对应的输入文字”这句话，当我输入“cong”时，输入法给我的候选字很多，

观测序列越长，模型能得到的信息越多，自然推断的准确性就越高。除了推断隐藏序列，HMM还可用作预测，即给定一段观测序列，预测下一个隐藏序列是什么。



基于GMM-HMM的语音应用

1）HMM

隐马尔科夫模型的缺点：

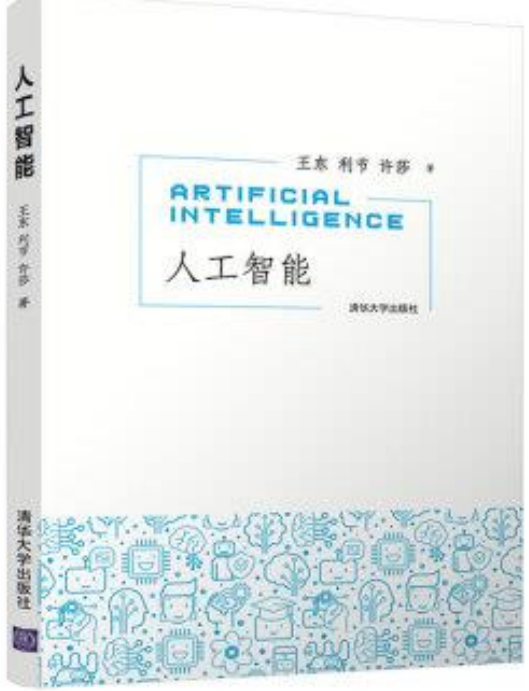
1、HMM只依赖于每一个状态和它对应的观察对象：

序列标注问题不仅和单个词相关，而且和观察序列的长度，单词的上下文，等等相关。

2、目标函数和预测目标函数不匹配：

HMM学到的是状态和观察序列的联合分布P(Y,X)，而预测问题中，我们需要的是条件概率P(Y|X)。

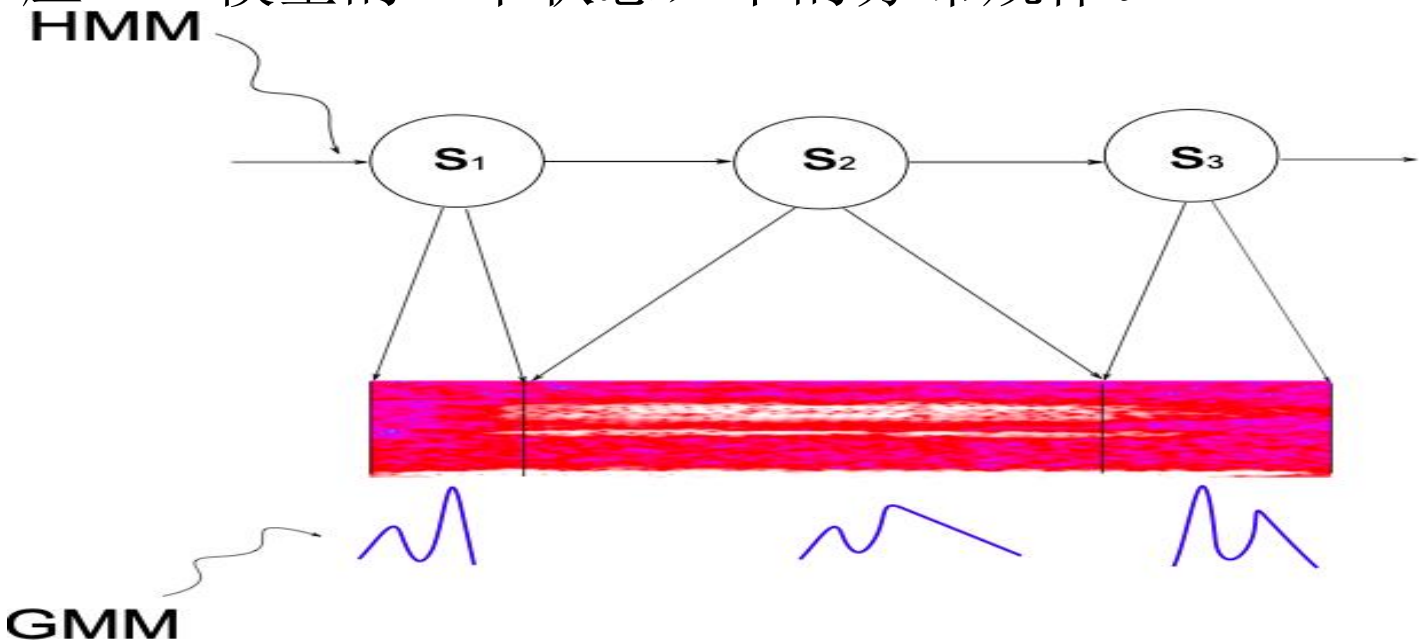
因此，深度学习引入语音识别的领域，并有较好的表现。

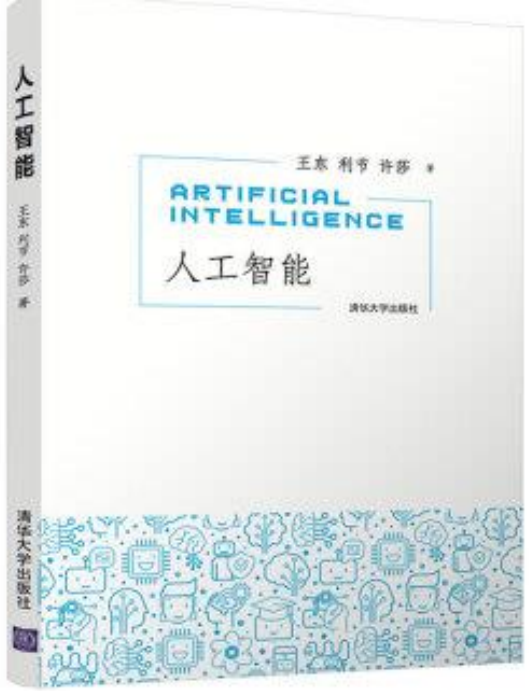


基于GMM-HMM的语音应用

2）GMM-HMM 声学模型

声学模型的作用是描述一个音素的发音过程。深度学习出现之前，GMM-HMM 是声学模型的主流。总体上来说，GMM-HMM 是一个概率模型。这一模型包括两个部分：描述发音动态特性的HMM 模型和描述短时静态特性的GMM 模型。动态特性是指语音信号在时间顺序上的发展演进过程；静态特性指语音信号在某个短时平稳状态（对应HMM 模型的一个状态）下的分布规律。



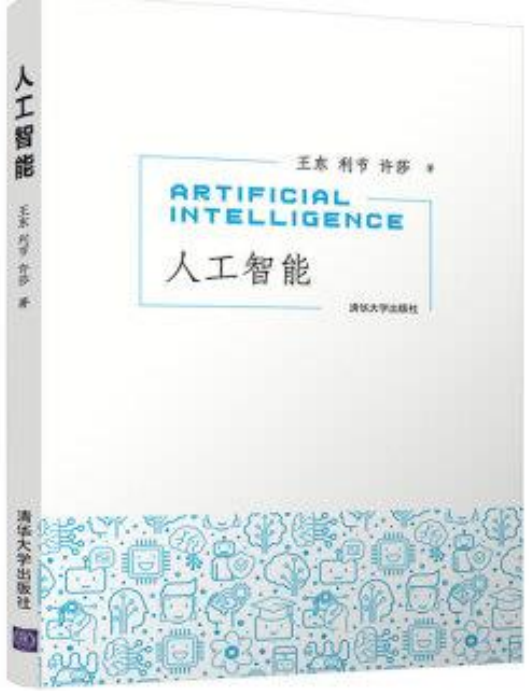


基于GMM-HMM的语音应用

3）N-gram 语言模型

语言模型的作用是描述语言中词与词的搭配规律。对语音识别而言，一种发音可能对应很多词，语言模型可以提供一种约束，帮助识别系统选择符合搭配规律的词。N 元文法（n-gram）是目前应用最广泛的语言模型。这一模型刻画的是给定n-1 个前序词，后接某个单词的概率。以3-gram 为例，给定前两个词为“我/吃/”，后接的词可以有很多选择，每种选择在实际中出现的概率各不相同，如：

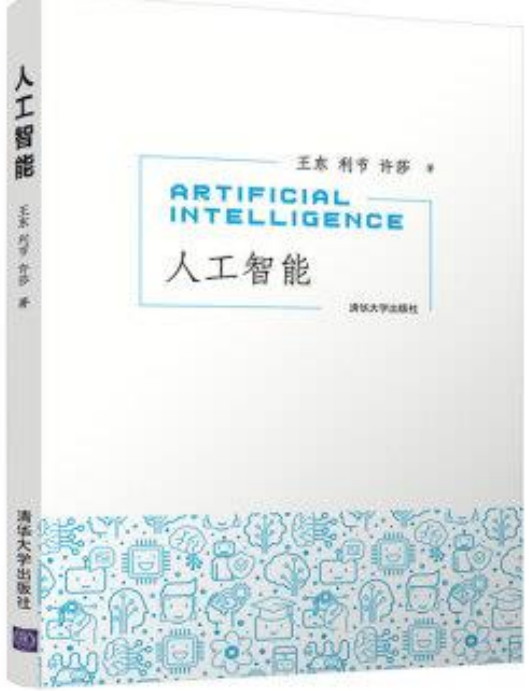
- 我/吃/水果0.1
- 我/吃/鱼0.2
- 我/吃/太阳0.0000003
- 我/吃/很0.000001
- 我/吃/去年0.00000004



基于GMM-HMM的语音应用

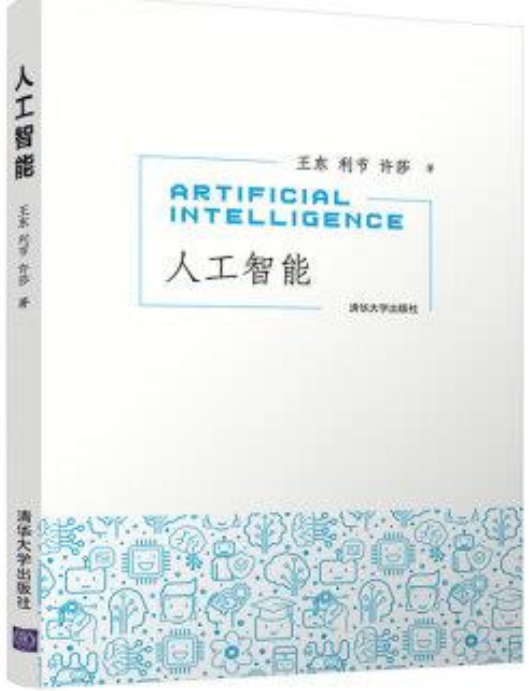
4）解码过程

现代语音识别系统的解码过程本质上是一个搜索过程：给定一段语音，在所有可能的句子中进行搜索，找到和该语音最匹配的句子。这里的匹配既需要考虑声学模型对语音信号的生成概率，也要考虑语言模型给出的词间搭配概率。考虑到单词组句的灵活性，这一搜索空间是非常巨大的。因此，一般采用一种称为剪枝的搜索策略：将语音特征向量依次输入解码器，每接收一个新的语音帧，解码器需要考虑加入一个新的音素或单词，这意味着搜索空间的扩展。为防止这种扩展失控，每次扩展后只保留当前匹配度最高的候选句子。这一过程称为剪枝搜索。



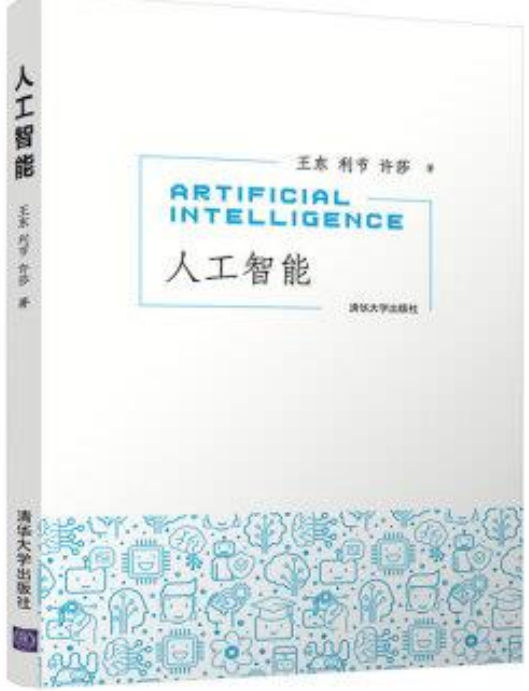
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用



基于深度学习的语音识别

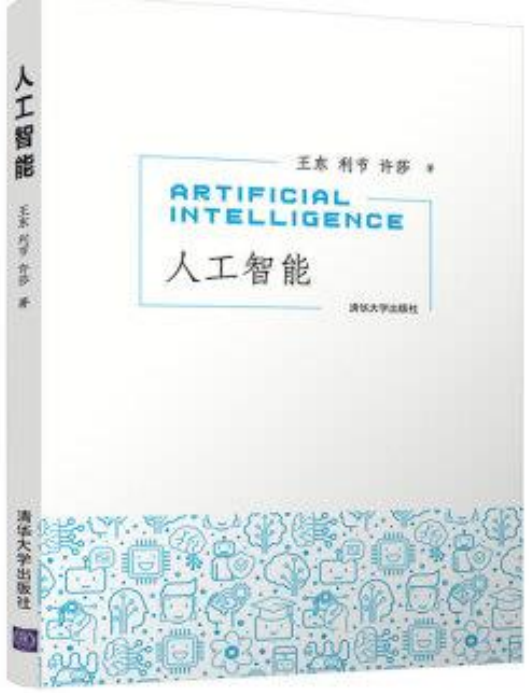
自2009 年以来，研究者将深度神经网络（DNN）引入到语音识别领域，试图突破概率模型的固有缺陷[62]。这一过程可分为三个阶段：第一步，利用DNN 来提取更有效的特征，以代替传统的MFCC；第二步，用DNN 代替GMM 计算HMM 中每个状态的概率；第三步，用递归神经网络（RNN）代替HMM 描述发音过程的动态变化。



基于深度学习的语音识别

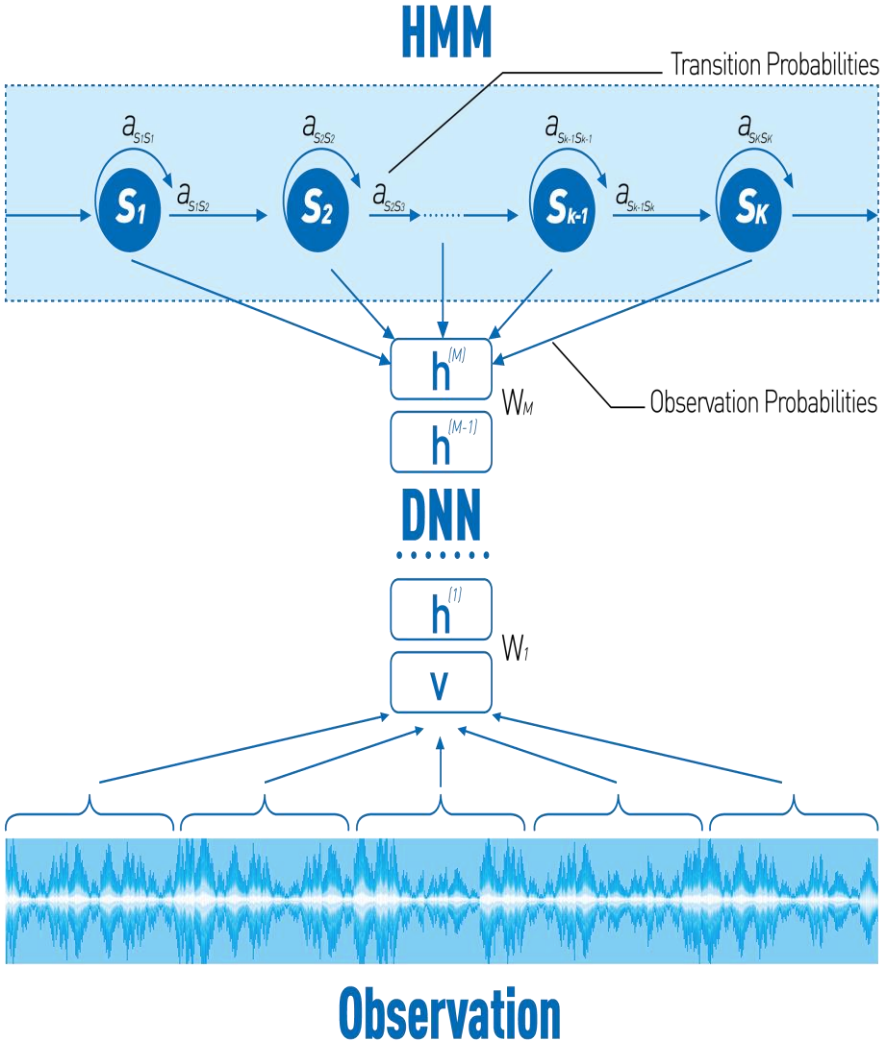
1）DNN 特征提取

MFCC 是统计模型时代的标准特征。MFCC 可以提取声道特性，对发音内容具有区分性。然而，这一特征受噪音干扰较大，在实际应用场景中区分度会下降。DNN 特征是从原始数据（可能是MFCC，也可能是其它初级特征）中通过学习自动发现的特征，具有更强的区分性。从中可以明显看到DNN 特征比MFCC 特征具有更强的音素区分性。

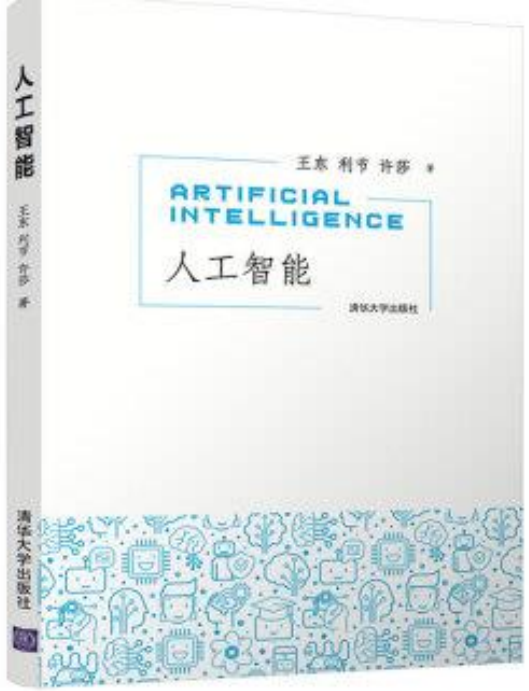


基于深度学习的语音识别

2）DNN 静态建模



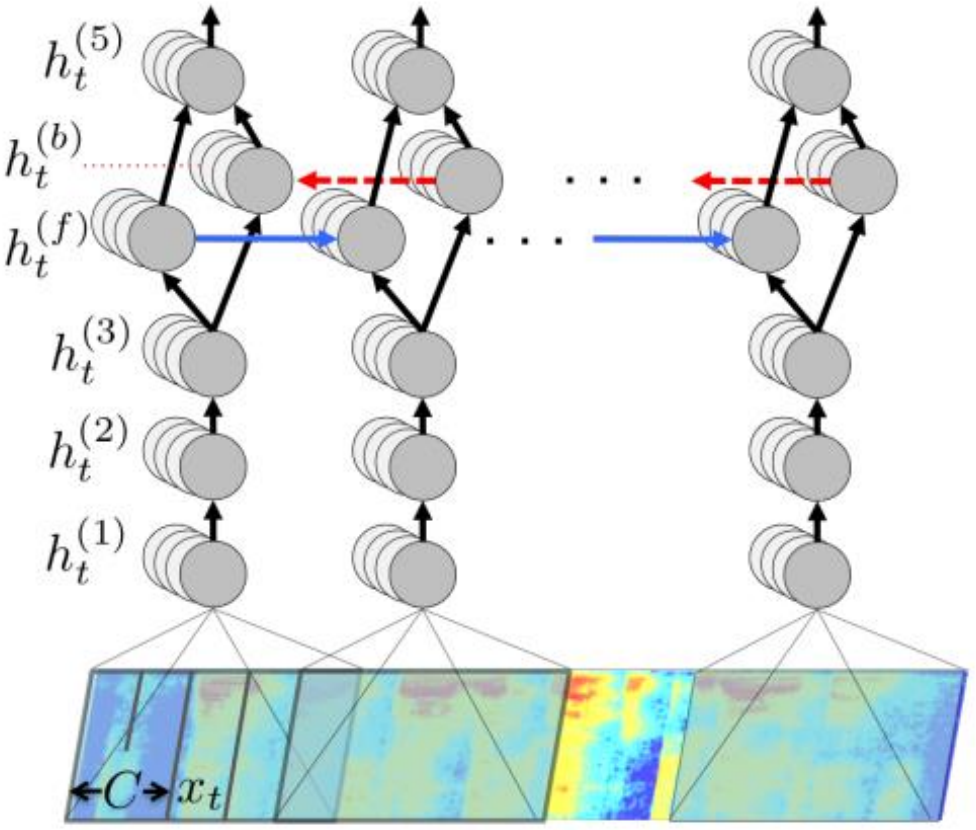
GMM 模型在静态建模方面有两个显著缺点，一是对复杂发音需要大量高斯分布才能实现较好的描述，效率不高；二是仅关注对发音本身的描述，不关心不同发音之间的区分性。DNN 模型恰好可以补足这些缺陷：它可以描述非常复杂的发音，且天然具有不同发音间的区分性。因此，研究者尝试用DNN 代替GMM 模型对HMM 中每个状态的特征分布进行建模。这一模型称为DNN-HMM 模型。



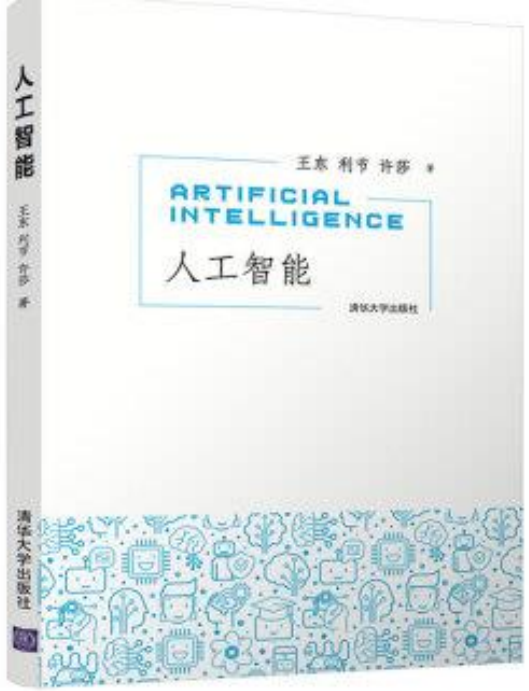
基于深度学习的语音识别

3）RNN 动态建模

HMM 模型将语音的发展过程简化为若干离散状态的序列，这显然不符合发音过程连续变动的实际情况。因此，研究者试图用一种连续动态模型取代HMM。

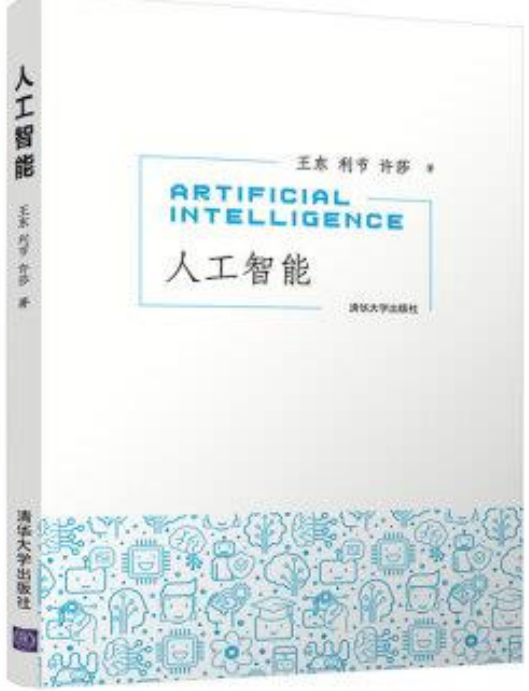


递归神经网络（RNN）正是这样一种模型。RNN 是神经网络的一种，由Jordan 等在80 年代提出。和标准前向神经网络（如多层感知器MLP）相比，RNN 在节点上加入了一个反馈连接，接收上一时刻的信息，因此具有记忆功能。和标准前向神经网络相比，RNN 可以学习时序上的相关性，因而常用于对序列进行建模。



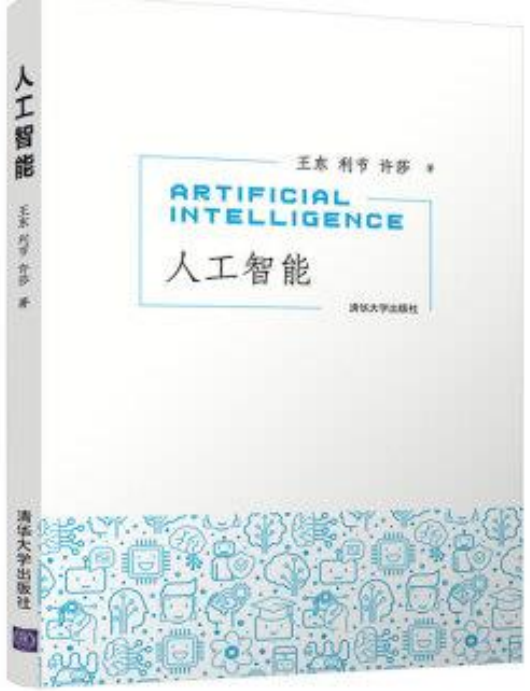
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用



说话人识别

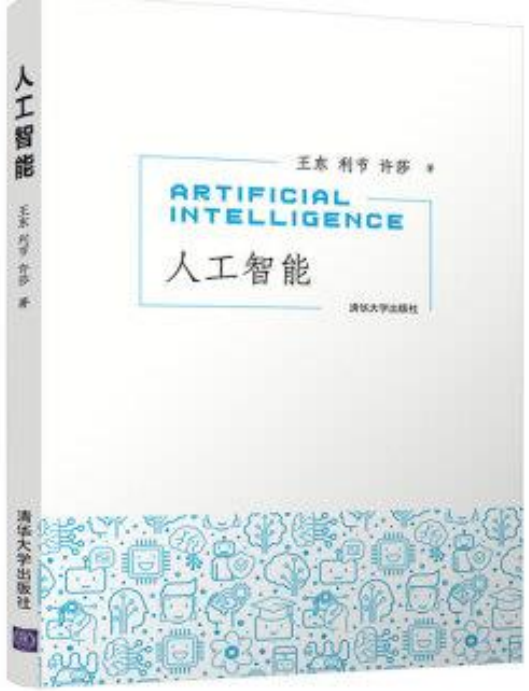
说话人识别是通过声音判断说话人身份的技术。和人脸识别一样，说话人识别也可以分为身份确认和身份辨认两种任务，前者对声音的身份进行判断，后者从数据库中找出发音人。和人脸识别相比，通过声音识别说话人有两个难点。第一，说话人信息并不是语音信号中的主要信息，容易受到其它信息的干扰，特别是发音内容和发音方式。而人脸图片的主要信息就是人的身份，其它因素的影响相对较小。第二，语音信号是时序信号，很难像图片那样互相对齐，导致模式匹配困难。对说话人识别的研究基本上是围绕以上两个困难（信息干扰和时序对齐）展开的。



基于深度学习的语音识别

1）传统GMM-UBM 系统

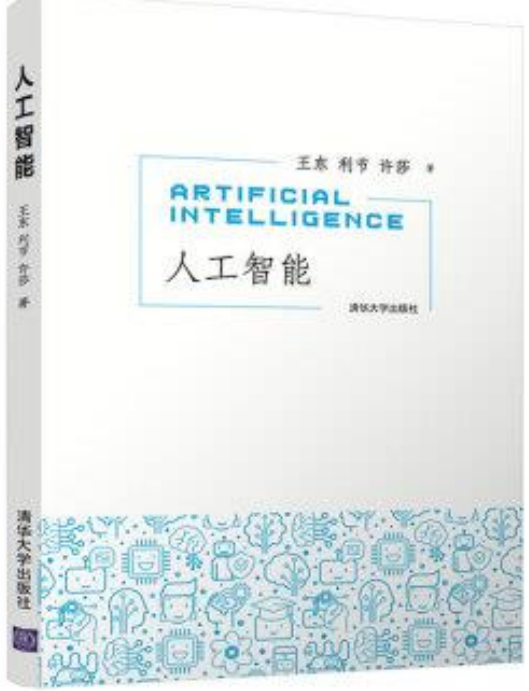
通用背景模型(Universal Background Model, UBM）将发音空间划分成若干子区域，每个子区域近似对应一种发音（如一个音素）并用一个高斯分布表示。因此，UBM 模型中包含多个高斯分布，是一个GMM 模型。对说话人建模时，将UBM 的每个高斯成分做少量偏移，得到新的GMM 模型即可代表该说话人。



基于深度学习的语音识别

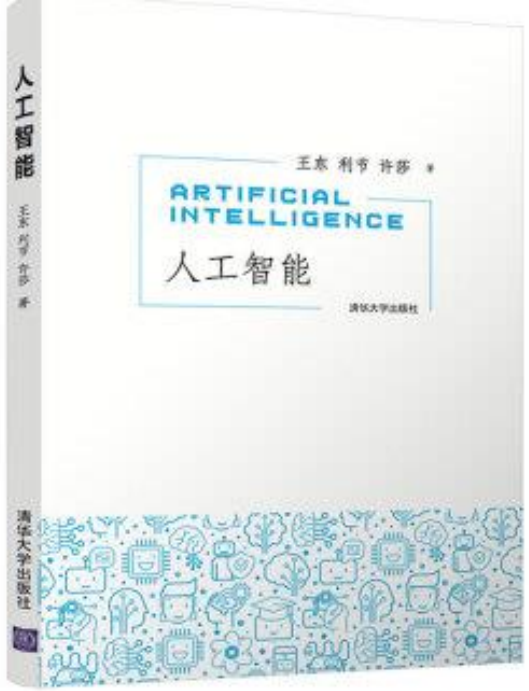
2）基于DNN 的说话人识别系统

网络的输入是原始语音帧，输出是训练集中的所有说话人。训练完成后，该网络即可实现从原始语音帧到说话人特征的逐层提取，而且越到后面，发音内容、信道等干扰因素被滤除得越干净，说话人信息也越显著。



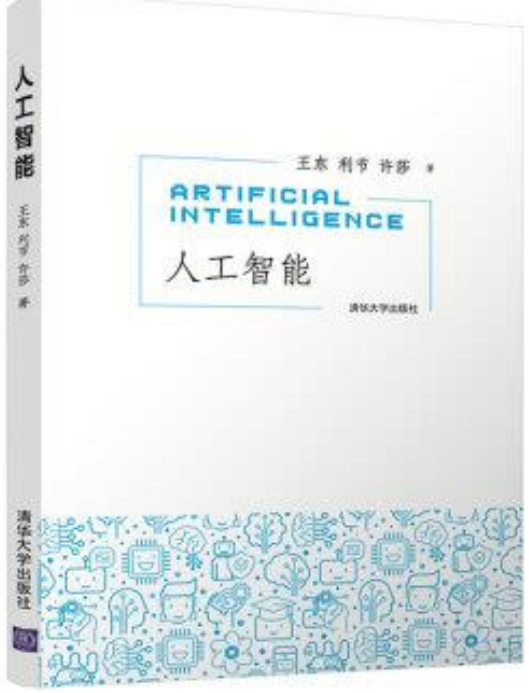
目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用



语音合成

- 语音合成是由文字生成声音的过程，通俗地说，就是让机器按人的指令发出声音。
- 计算机发明以后，语音合成技术开始快速发展。按时间顺序，语音合成方法可归纳为四种：参数合成、拼接合成、统计模型合成和神经模型合成。



语音合成

1）参数合成

1930 年，贝尔实验室发明了声码器，将声音分解成声带激励和声道调制两部分（即本章开始提到的激励-调制模型）。

DECtalk lets micros read messages over phones

BY PEGGY ZIENTARA
Senior Editor

The use of the telephone as a "universal terminal" came one step closer to reality when Digital Equipment Corporation (DEC) introduced a voice synthesizer unit that can "read" aloud ASCII text messages.

DECtalk, which the company says will be available in March at a price of $4000, allows you to call in to a personal computer from anywhere in the world by touch-tone telephone (including pay phones), to hear electronic mail messages stored in memory.

DECtalk can also "tell" you what you just typed into memory on the computer keyboard if you simply press a punctuation mark key and the Return key.

Business applications are expected to include access to company data bases by traveling sales representatives; access by lawyers to public and private computer files during trial breaks; direct customer transactions with banks and stock brokerage firms; direct catalog and retail sales; and access to transportation schedules, according to Ed Kramer, vice-president of corporate marketing for DEC.

Other applications will include aid for the speech-impaired (now in place at

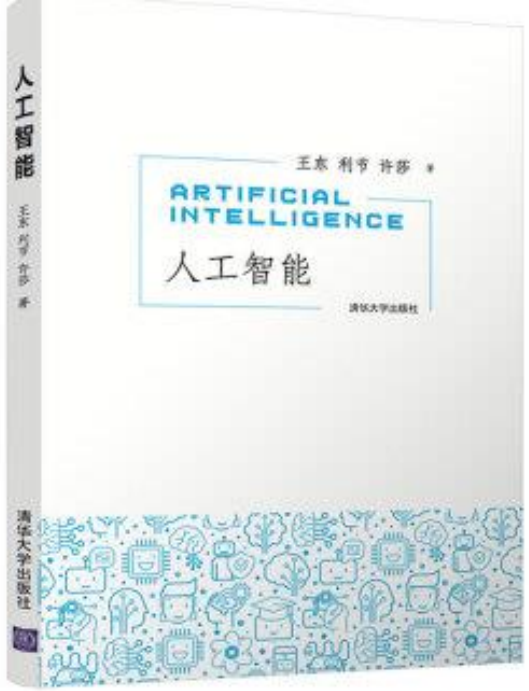
via any touch-tone phone.

The DECtalk hardware, which fits easily under a 12-inch monitor, will feature eight different types of voices, including that of an old man, a deep-voiced man, a typical middle-aged man, two different women and a child. The unit attaches to a personal computer via an RS-232C serial interface.

DEC claims to have licked the traditional voice technology trade-offs of having either good voices with limited vocabularies or unlimited vocabularies with poor voices. DEC's unit has a "virtually unlimited vocabulary," Kramer said, and the speaking rate can be adjusted from 120 to 350 words per minute. Tone and modulation can also be regulated.

Heuristics — interpreting word context to improve pronunciation — and a computer model of the human vocal tract allow DECtalk to produce high-quality voices, in contrast with traditional stored digitized or synthesized voice technology, Kramer explained.

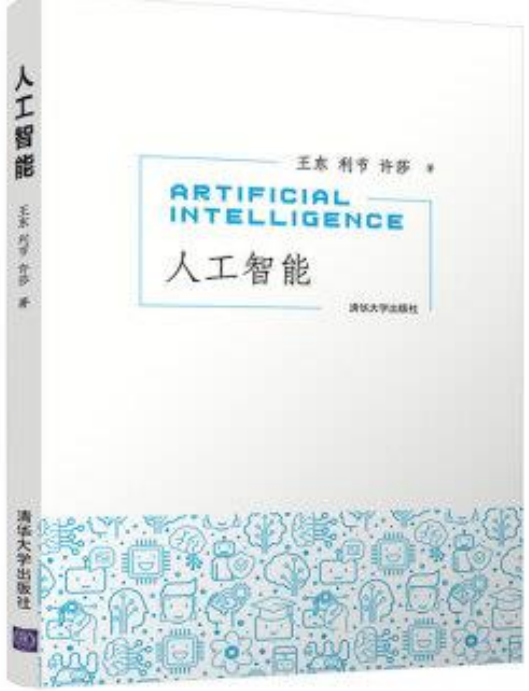
The unit accepts ASCII text, which it applies to an internal dictionary and a set of letter-to-sound rules indicating pronunciation and context interpretation to



语音合成

2）拼接合成

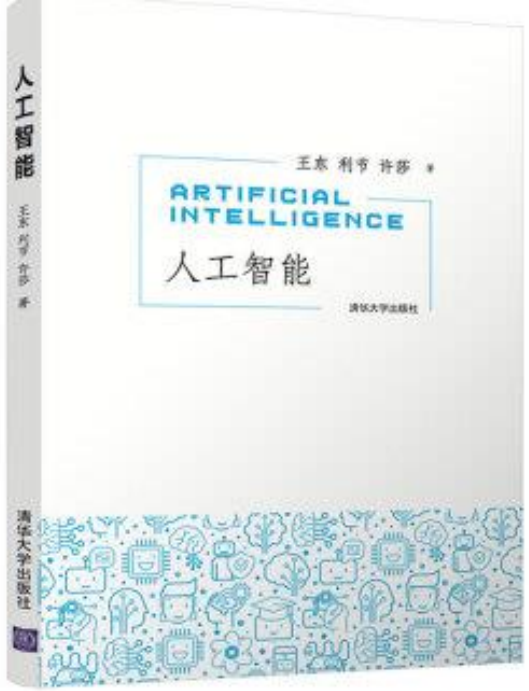
将事先录好的语音切分成发音片段（一般为音素），在合成时从这些片段中选择合适的候选进行拼接组成句子。



语音合成

3）统计模型合成

基于隐马尔可夫模型（HMM）的统计模型方法是这一时期的主流。这种方法对每个发音单元建立一个HMM 模型，在合成时将句子中所有音素的HMM 模型拼接起来形成一个组合模型，由该模型生成最匹配的语音（可参考语音识别中由GMM-HMM模型生成语音过程的说明）。

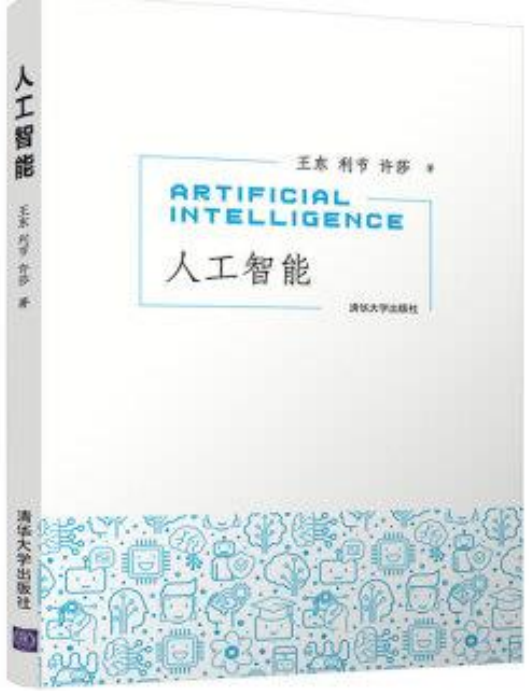


语音合成

4）神经模型合成

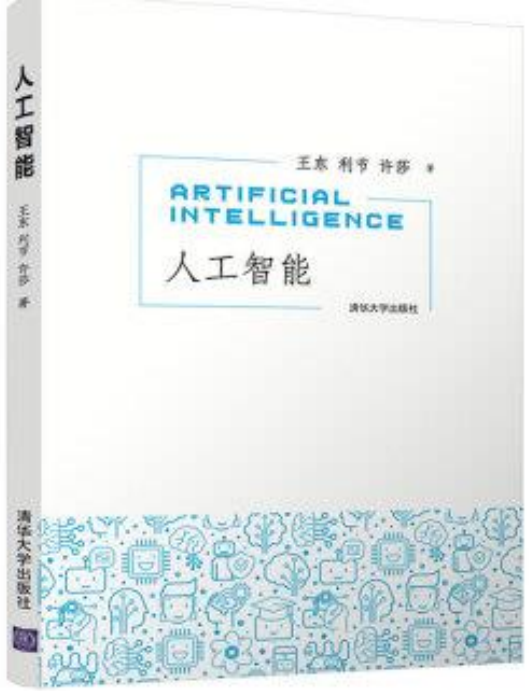
基于深度神经网络（DNN）的语音合成方法由香港中文大学、微软、Google 于 2013 年提出。用神经网络取代HMM 模型来预测每一帧语音的激励和调制信

号，再通过声码器合成自然语音。



目 录

- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用



智能家居控制系统

个性化配音服务

物联网输入法

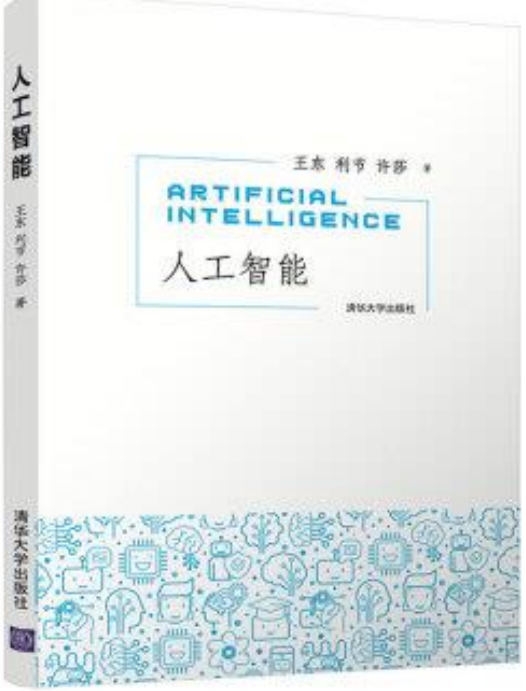
语音

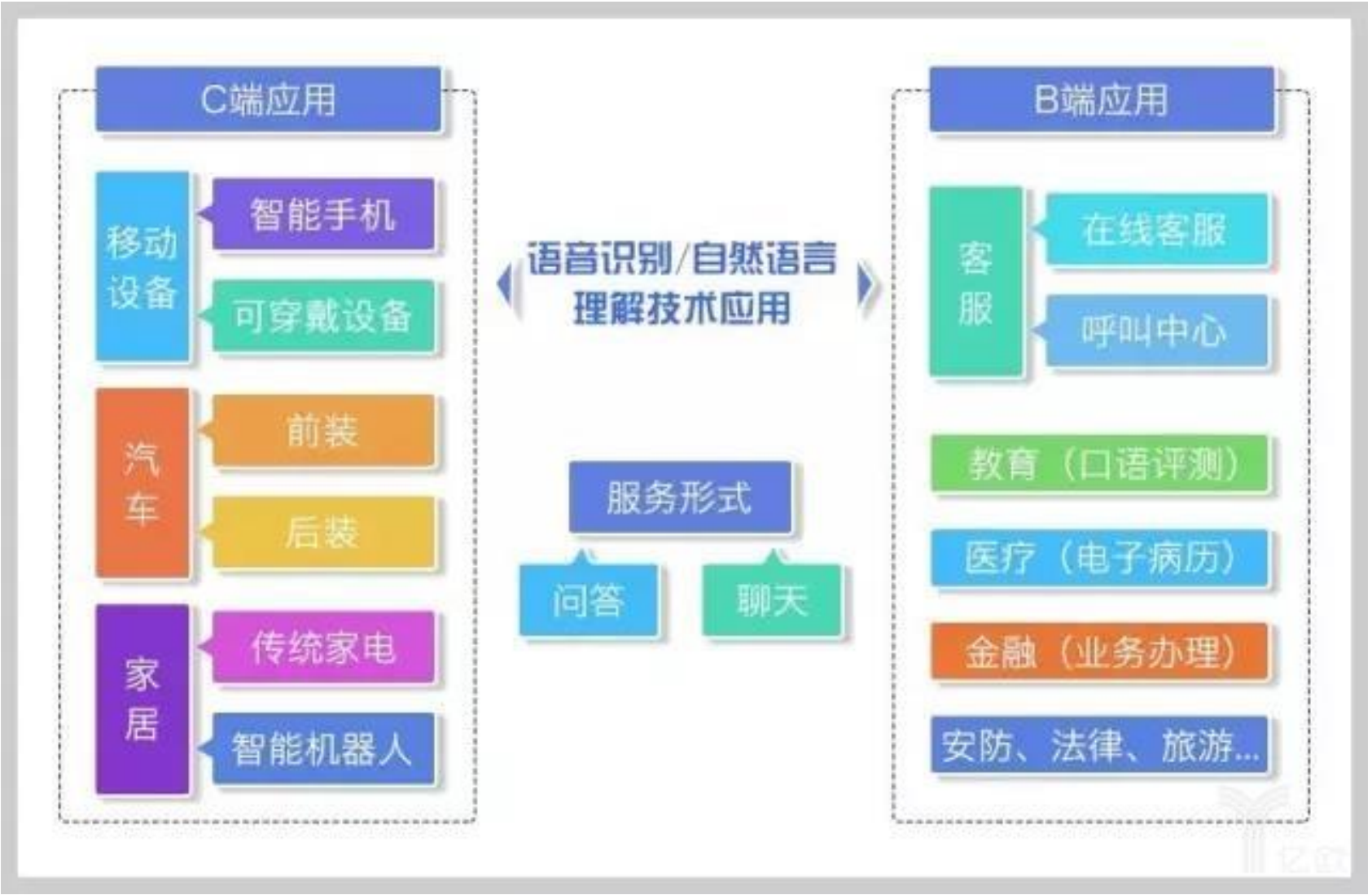
个性化语音合成

中英互译（其他语言）

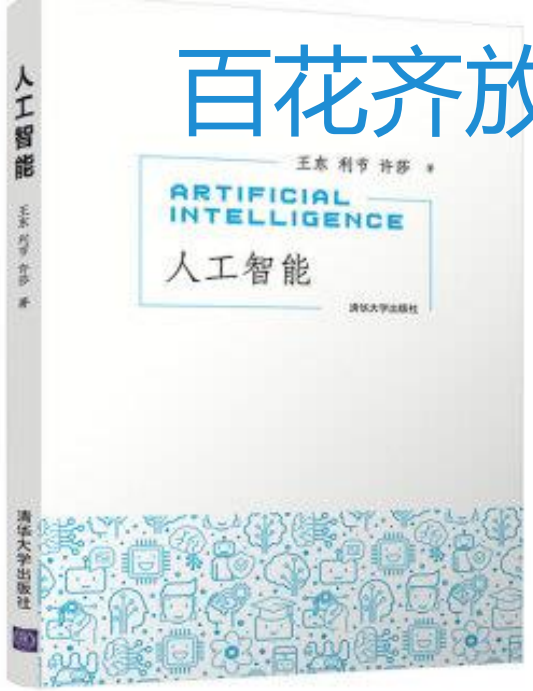
智能客服机器人

汽车智能车载系统


人工智能
王东 利亨 许莎 著
ARTIFICIAL INTELLIGENCE
人工智能
清华大学出版社

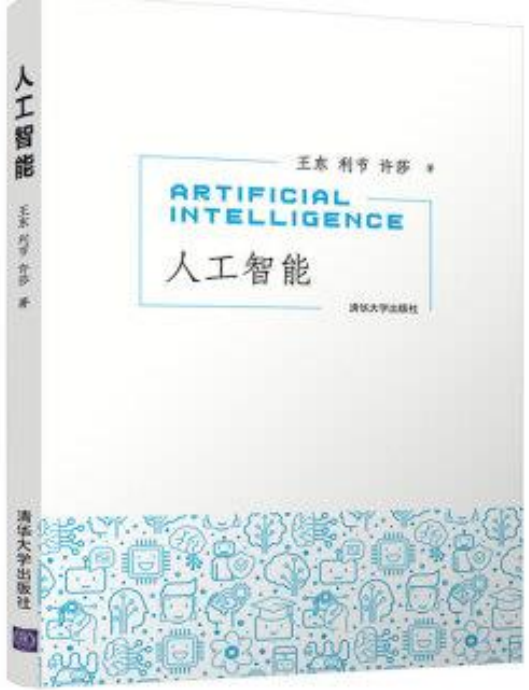

C端应用
移动设备
智能手机
可穿戴设备
汽车
前装
后装
家居
传统家电
智能机器人
语音识别/自然语言理解技术应用
服务形式
问答
聊天
B端应用
客服
在线客服
呼叫中心
教育（口语评测）
医疗（电子病历）
金融（业务办理）
安防、法律、旅游...

百花齐放



Google 科大讯飞 iFLYTEK Baidu 百度 amazon.com

百度语音开放平台：每日在线语音识别调用**1亿4千万**次（2016.11）。



The end !