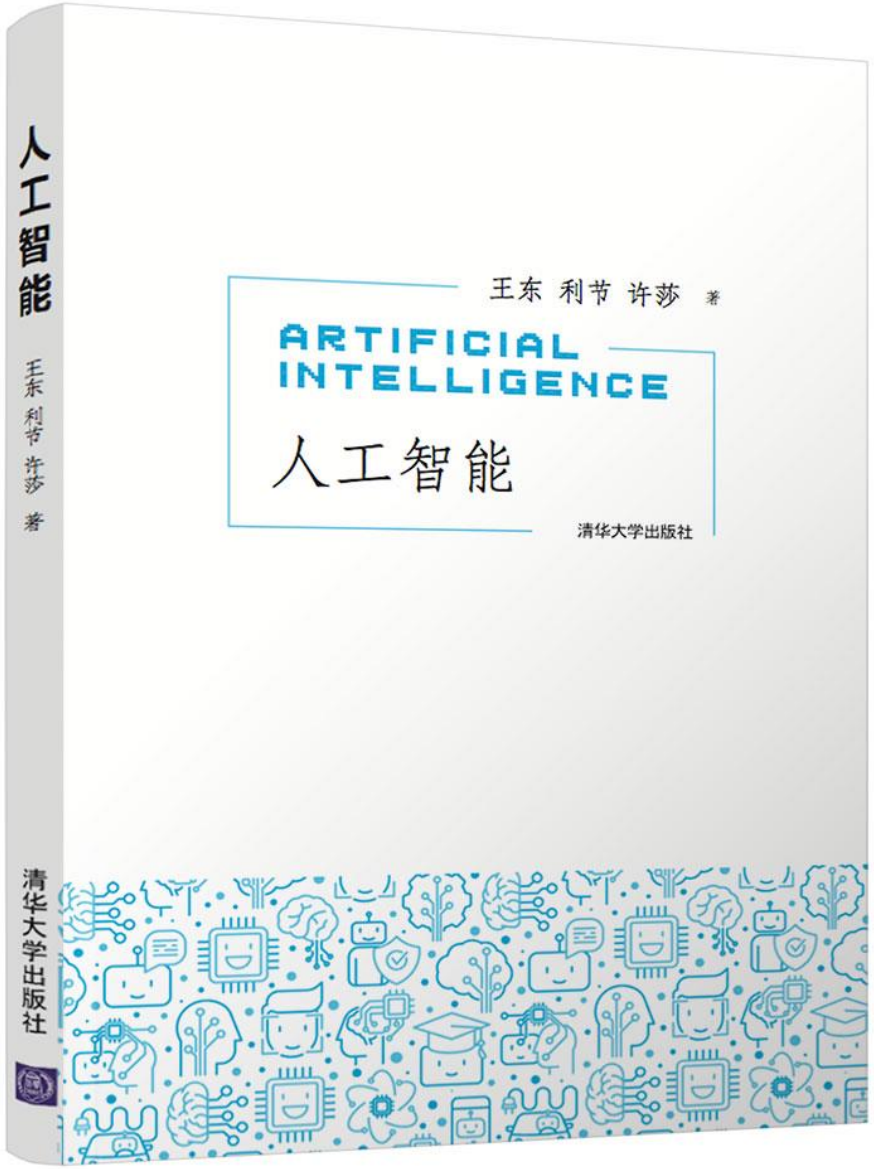
人工智能

王东 & 利节 & 许莎



- 教材名称：人工智能
- 出版社：清华大学出版社
- 作者：王东 利节 许莎

本课程应具备的能力：

- Linux下的基本操作
- python语言基础知识

资源下载：

http://aibook.cslt.org/

倾听你的声音

汤志远

授课提示

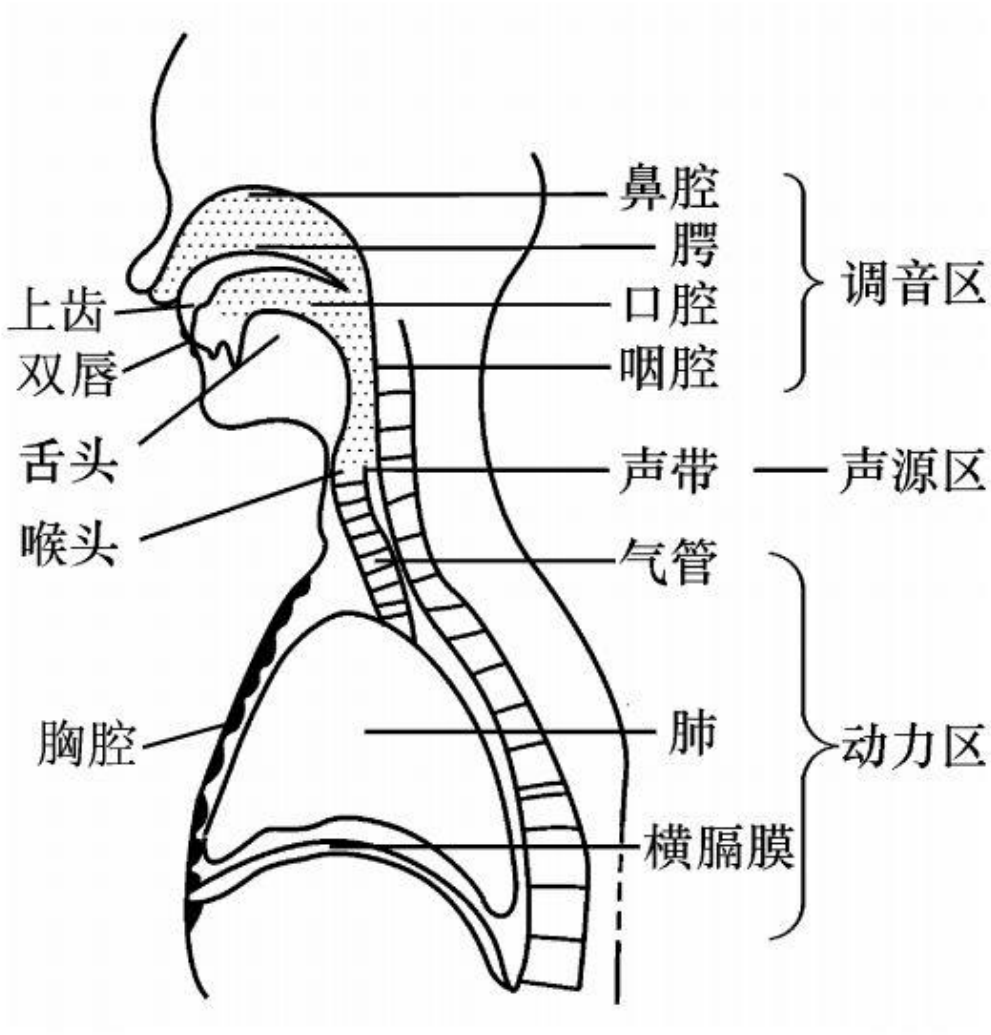
1. 本课件70页，建议授课时间180分钟，每页平均2.5分钟。重点可放在“基于深度学习的语音识别”、“说话人识别”和“语音合成”三部分内容，以提高学生对常用语音技术相关概念的了解。
2. 若授课时间为90分钟，请去掉非课本内容，对“附注”中所提内容尽量简化，并对具体的技术细节进行缩减。
3. 课件内的视频具有较好的展示效果，可视授课时间长短决定取舍。
4. 建议教授过程中关于“语音技术给生活带来了哪些影响”、“未来的语音技术应用可能有哪些”等问题进行讨论，激发学生思考。

语音助手--谁是世界上最美的人?

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

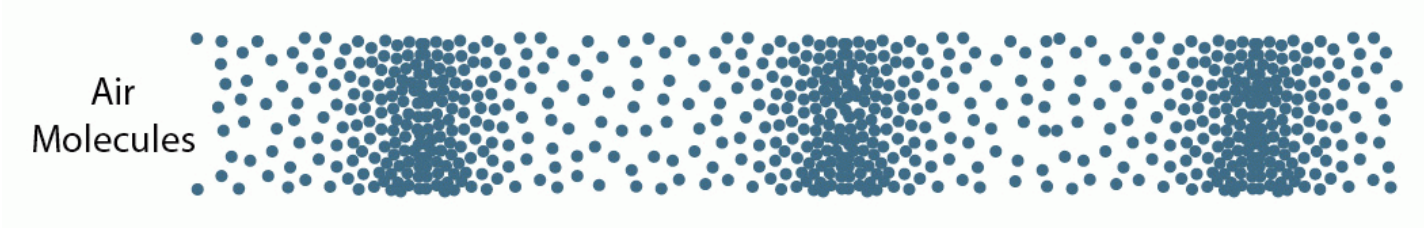
语音的产生与感知



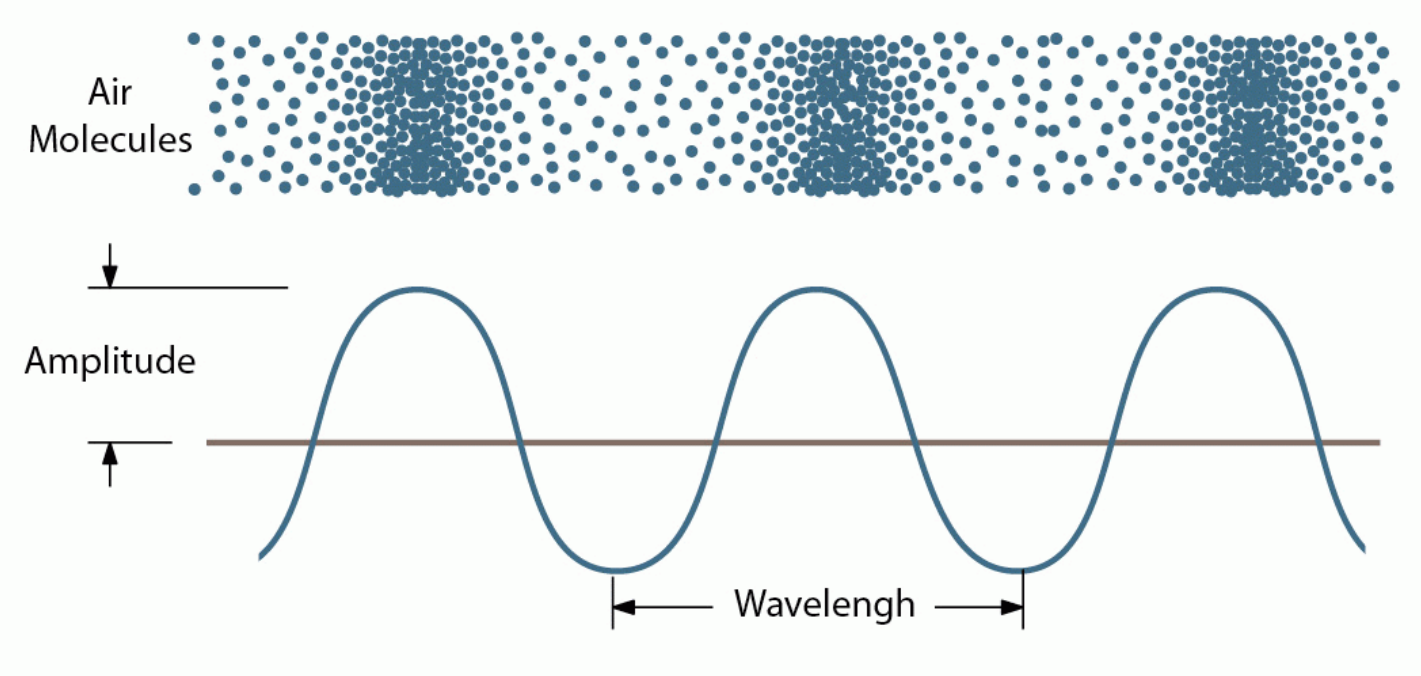
发音器官示意图

肺部产生气流动力，经过气管引起声带振动形成声源，最后经过咽腔、口腔、鼻腔等区域调整后发出最终的语音。不同于其他动物通过相同机制发出的声音，人类语音承载着丰富的语义，是人类的显著特征之一。

语音的产生与感知

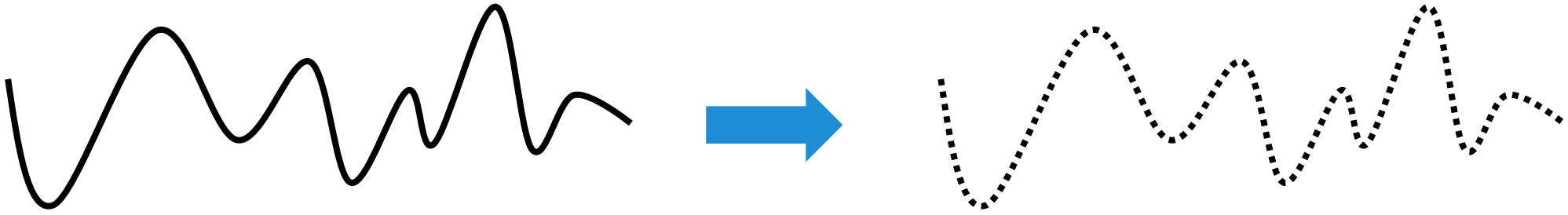


语音的产生与感知



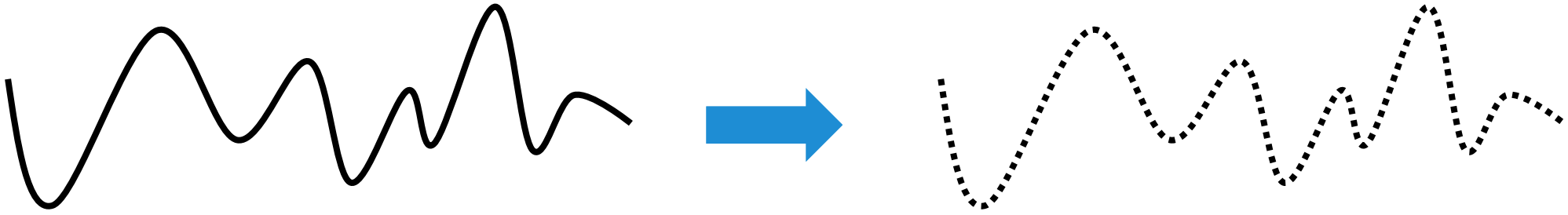
频率->音调，振幅->响度，二者->音色。
以不同振幅和频率敲打耳膜，人耳听到不同的声音。

计算机眼中的语音



模拟信号　　数字信号

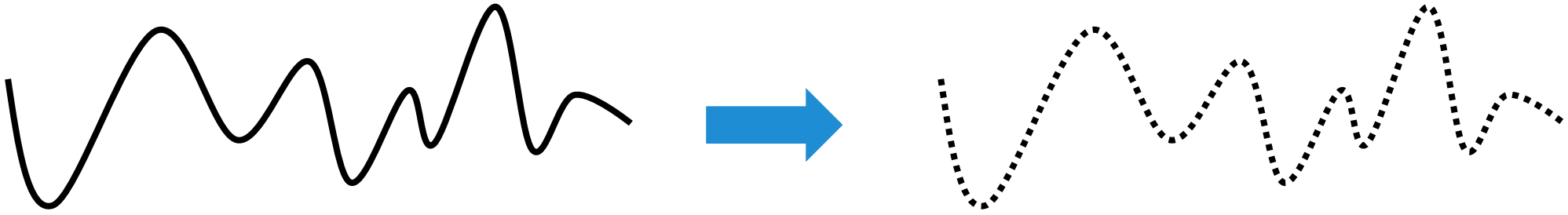
计算机眼中的语音



模拟信号　　数字信号

魔术转换

计算机眼中的语音



模拟信号　　数字信号

模数转换

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 +0 dB 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

dB -3 -6 -12 -18 -∞ -18 -12 -6 -3

L

R

1:20.833

ts Rack 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

ts Rack 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

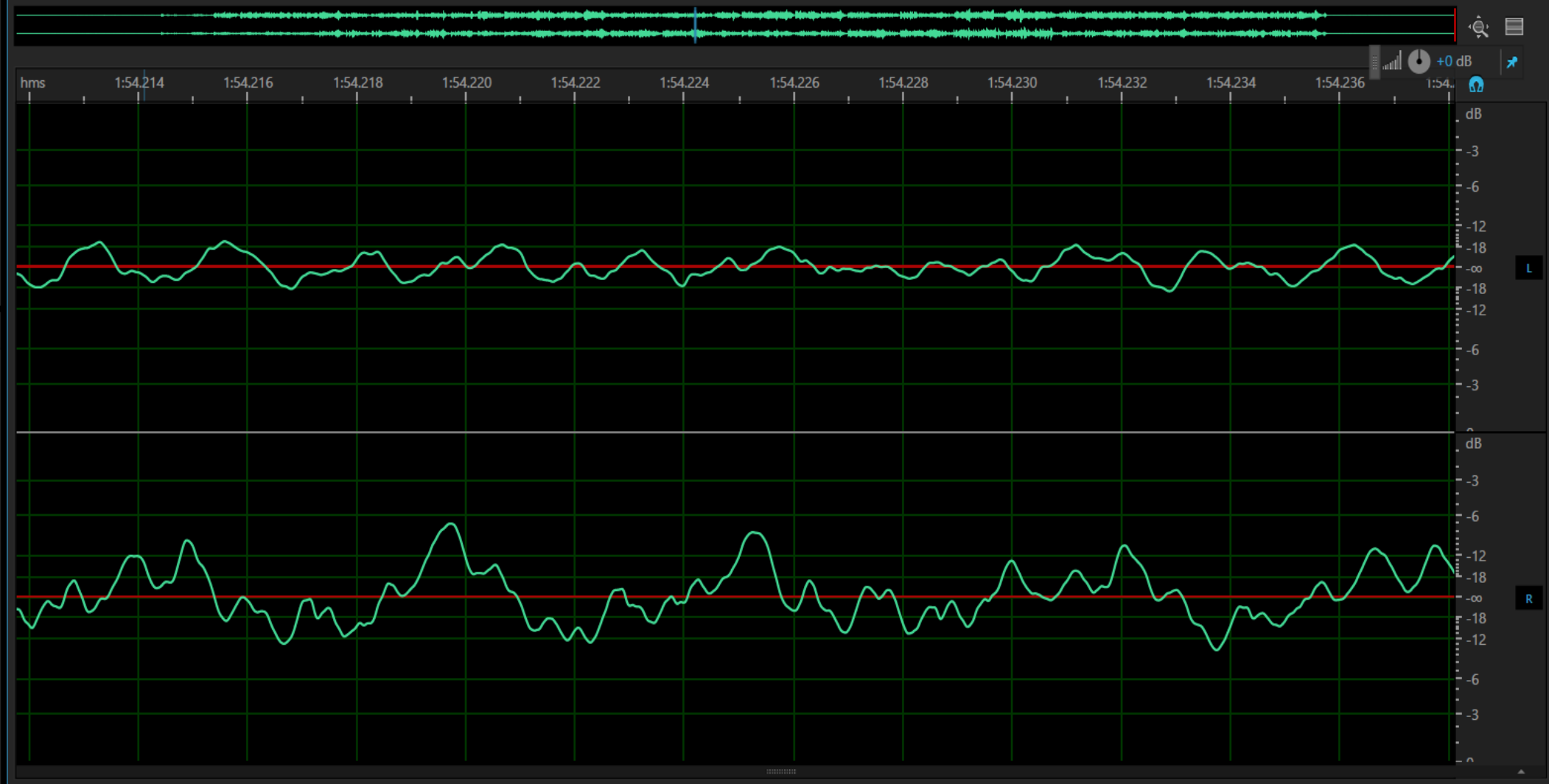
Time Display

History Video

Open

0 Undo

Editor: 萤火虫-牛车上的阳光.mp3 Mixer



4:01.972

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 -∞ -18 -12 -6 -3 L

dB -3 -6 -12 -18 -∞ -18 -12 -6 -3 R

4:01.972

ts Rack 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

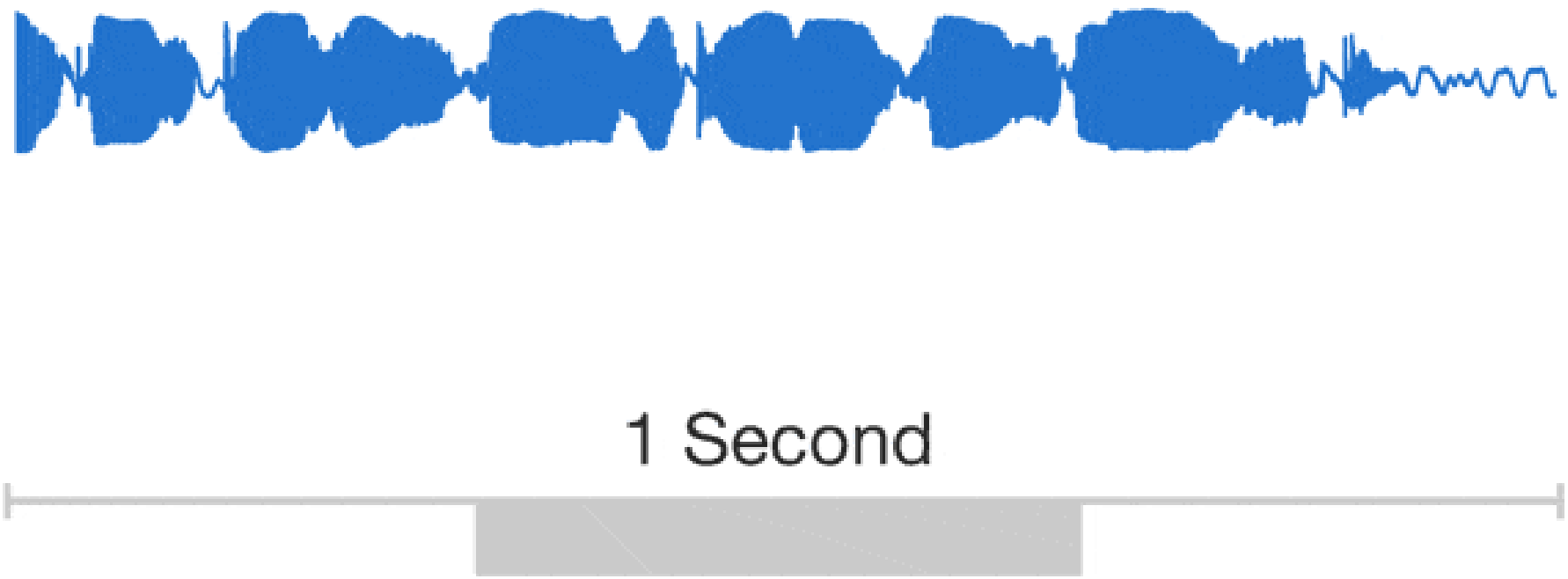
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0

Selection/View

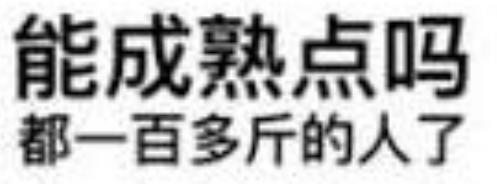
	Start	End	Duration
Selection	4:01.972	4:01.972	0:00.000
View	1:54.224	1:54.225	0:00.000

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



语音中的信息


能成熟点吗
都一百多斤的人了

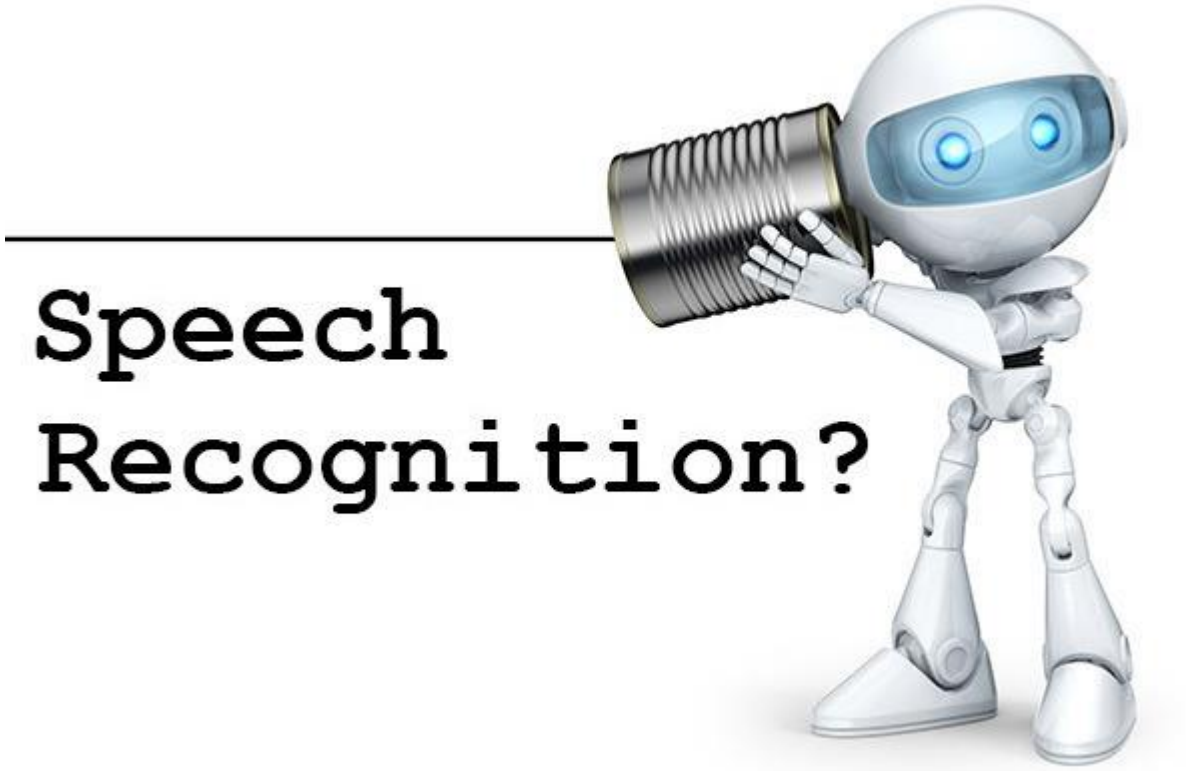
English
Spanish
Mandarin
French
Arabic
language
Japanese
Hindi
Thai
Korean
Portugese
Russian
Polish
German
Vietnamese
Italian
Cantonese
Czech
Greek
Hungarian
Ukrainian
Romanian
Persian
Swedish
Dutch
Bengali
Mandarin

蓝瘦

目标

- 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

语音识别是什么



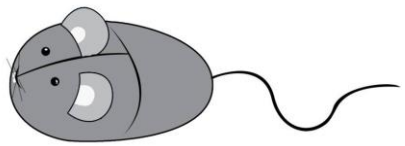
近在眼前的语音识别

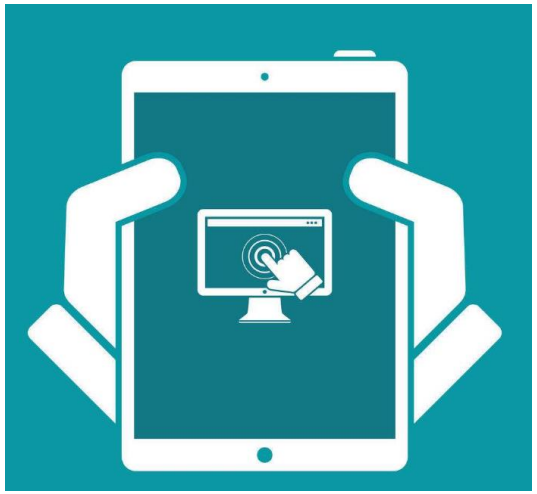
17:38
0.02K/s
16%
文件传输助手
使用听筒模式
收藏
提醒
转换为文字（仅普通话）
删除
更多
按住 说话

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

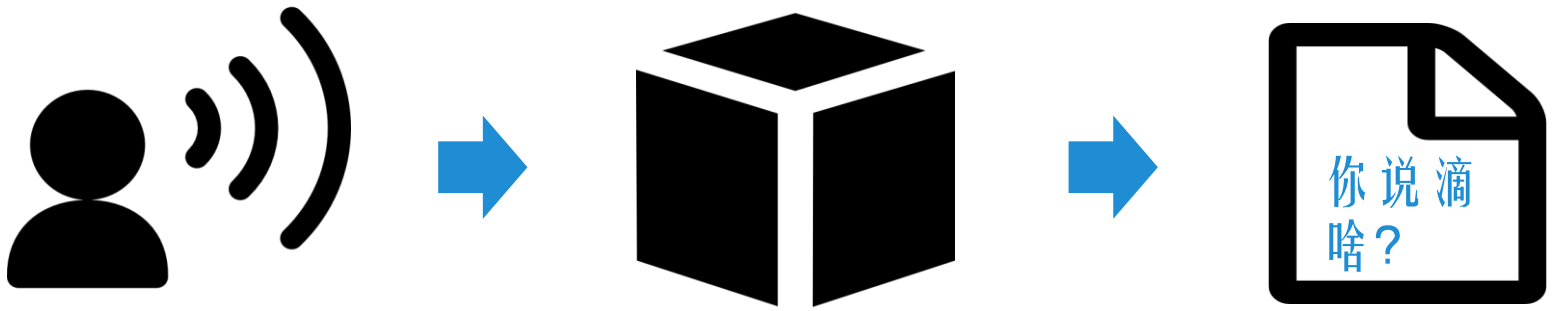
语音识别的重要性

人机交互方式?





语音识别是什么



Automatic Speech Recognition (ASR)

语音识别的历史

语音识别发展史	主要时间段及技术方法
第一阶段	20 世纪 30 年代至 50 年代，针对特定语音或小词表的孤立词设计专门的语音识别算法或装置。
第二阶段	20 世纪 50 年代至 60 年代，基于声学（Acoustics）和语音学（Phonetics）知识设计语音识别系统，使用模板匹配（Template Matching）方法来识别音位、音素和数字。
第三阶段	20 世纪 60 年代至 80 年代，使用模式识别（Pattern Recognition）方法进行中小规模的孤立词或连续词串语音识别。
第四阶段	20 世纪 80 年代以后，隐马尔科夫模型（Hidden Markov Model，HMM）用以处理语音中的动态问题，并用于连续语音识别系统设计，基于统计模型（Statistical Model）的方法慢慢确立成为语音识别主流框架；21 世纪初期，机器学习的发展，特别是深度学习的崛起，大大促进了大规模连续语音识别系统的性能提升和实用化。

语音识别的历史

1784

Wolfgang von Kempelen creates the **Acoustic-Mechanical Speech Machine** in Vienna

1879

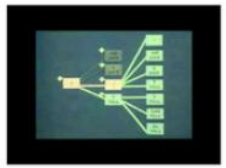
Thomas Edison invents the first dictation machine

1952

Bell Labs releases **Audrey**, capable of recognizing spoken digits with **90% accuracy** - but only when spoken by its inventor

1962

IBM Shoebox can understand 16 English words



1971

Harpy, created at Carnegie Mellon University, can comprehend **1,011 words** - and some phrases

1986

IBM Tangora, using the Hidden Markov Model, predicts upcoming phonemes in speech

2006

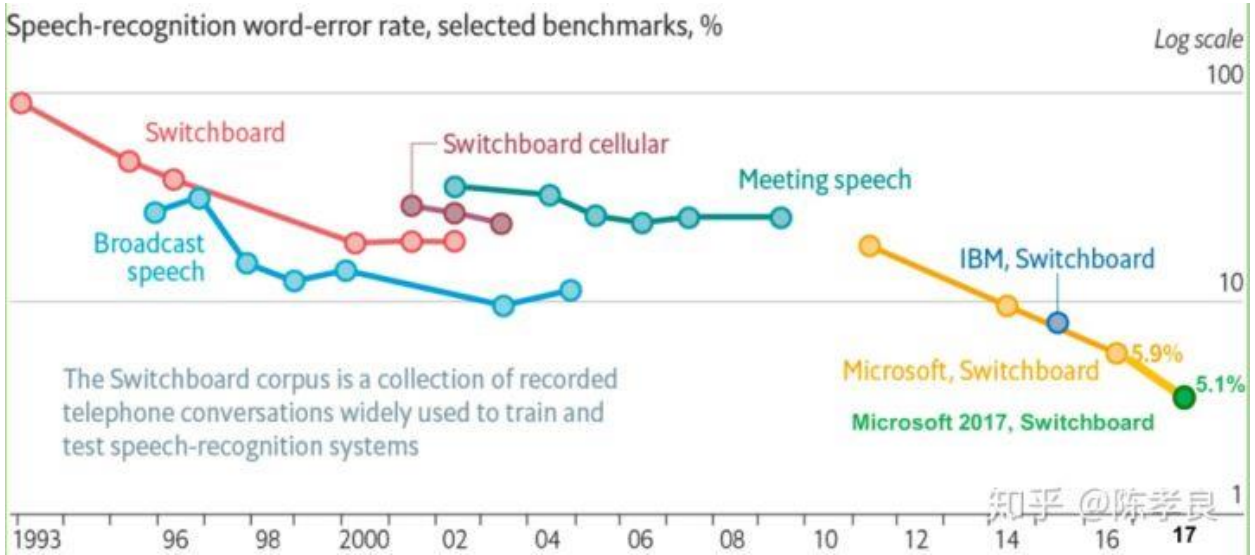
The National Security Agency (NSA) starts using speech recognition to isolate key words in recorded speech

2008

Google launches a voice search app, bringing speech recognition to mobile devices

2011

Apple announces Siri, ushering in the age of the voice-enabled digital assistant



知乎 @陈孝良

语音助手

Siri

Google ASSISTANT

Cortana

alexa

智能音箱

喜马拉雅小雅

若琪

天猫精灵

小米AI音箱

叮咚音箱

出门问问

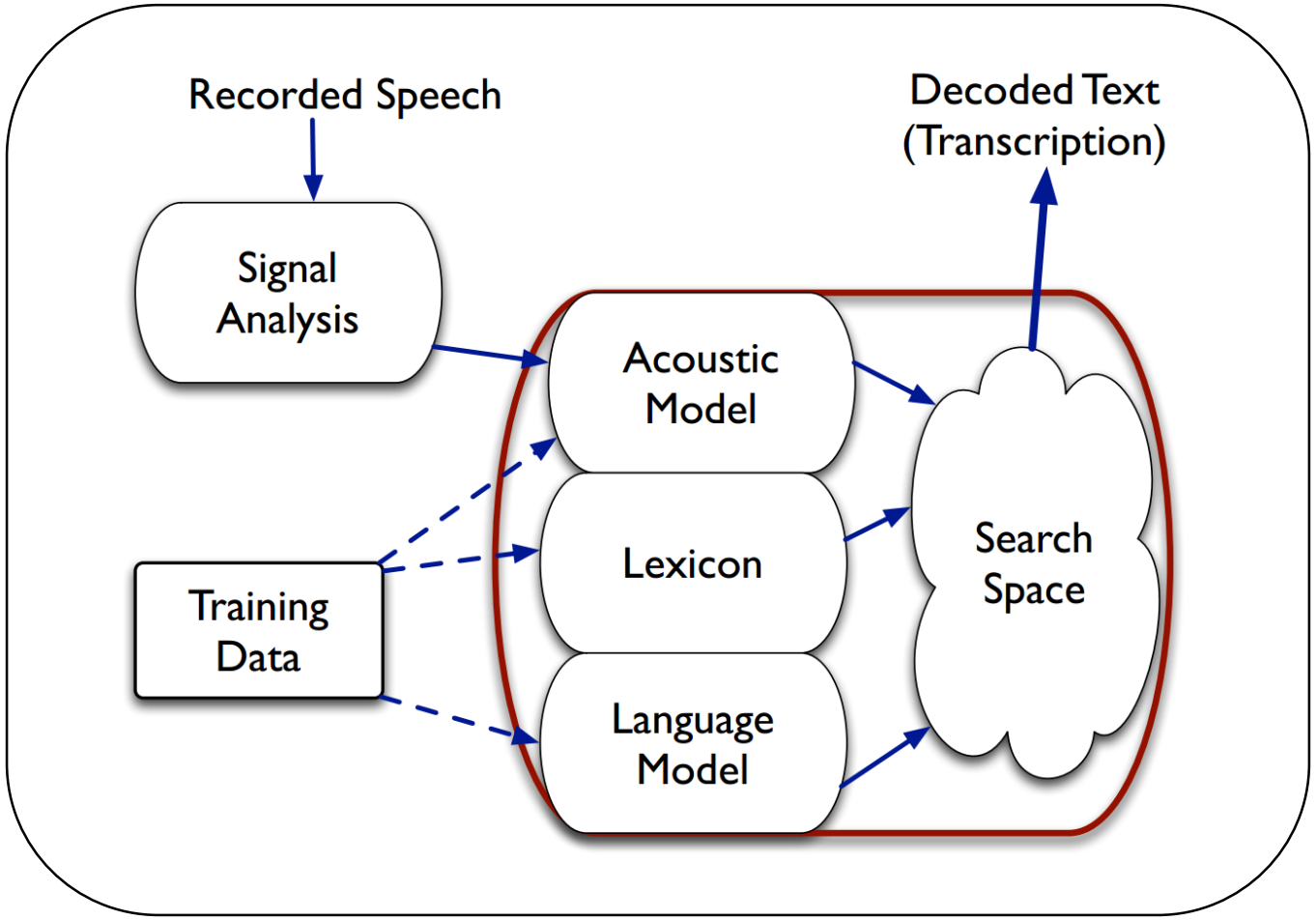
亚马逊Echo

Google Home

同声传译

同声传译使用了哪些技术?

研究者眼中的语音识别



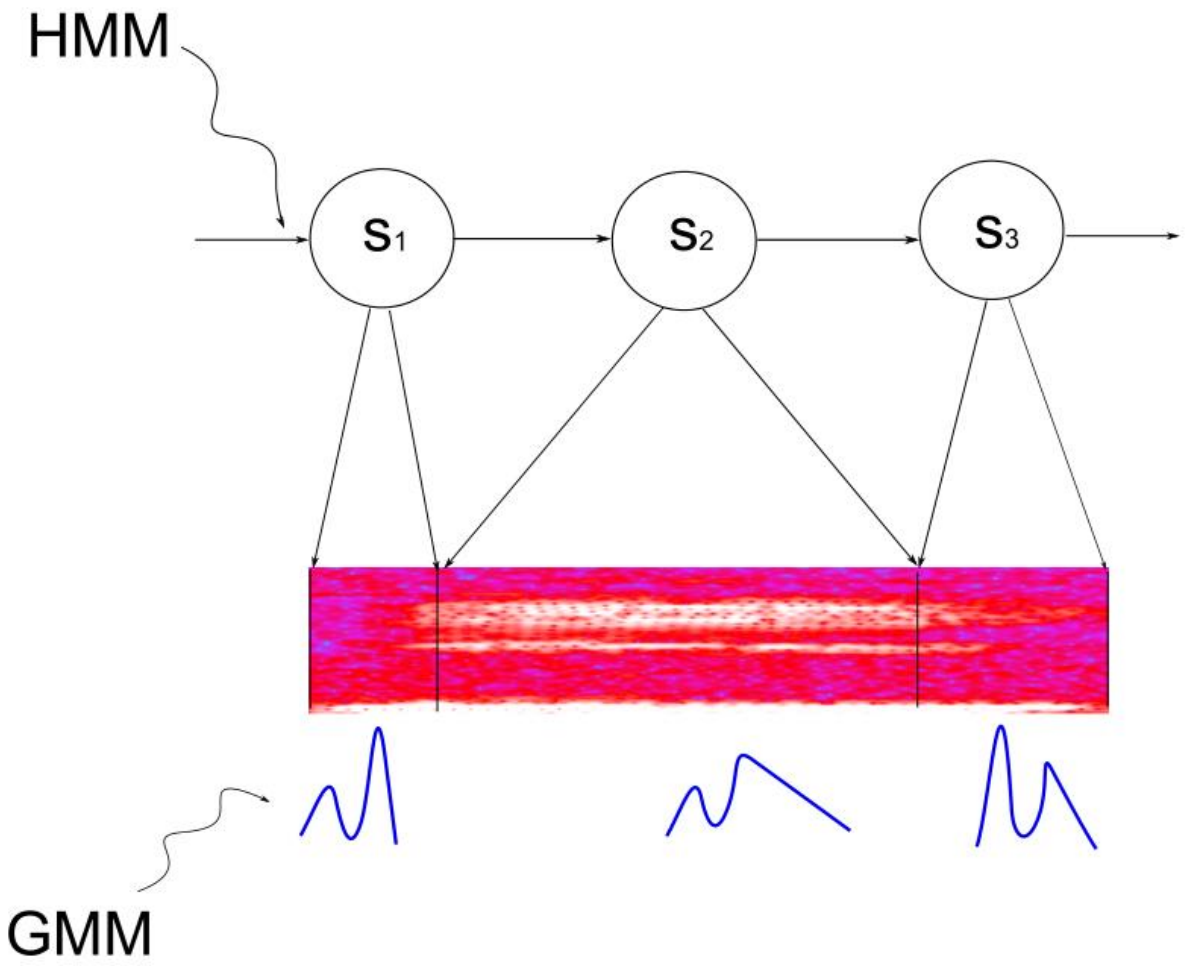
音频 ➡ 声学特征 ➡ 音素 ➡ 字/词 ➡ 句子

声学模型　　语言模型

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

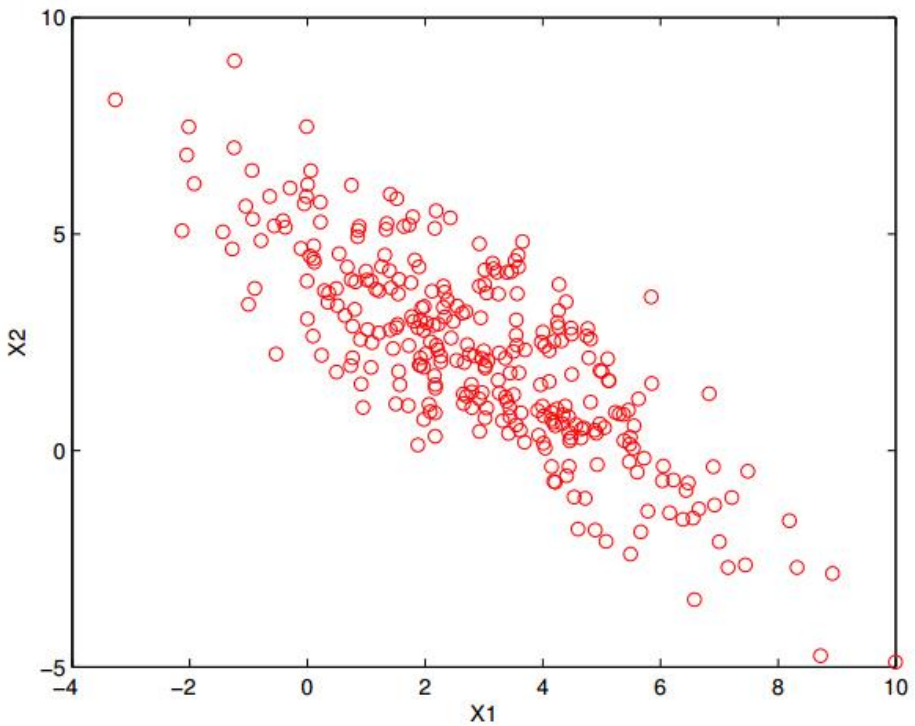
GMM–HMM

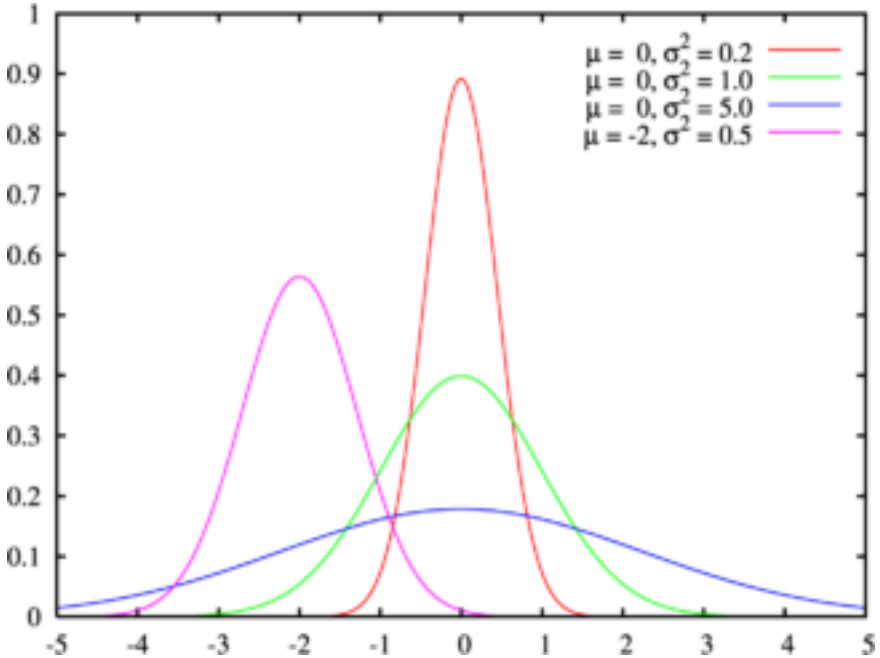


高斯模型

高斯分布, Gaussian Distribution， 也叫自然分布或正态分布，Natural Distribution。

正态分布有一个非常重要的性质：在特定条件下，大量统计独立的随机变量的和的分布趋于正态分布，这就是中心极限定理。

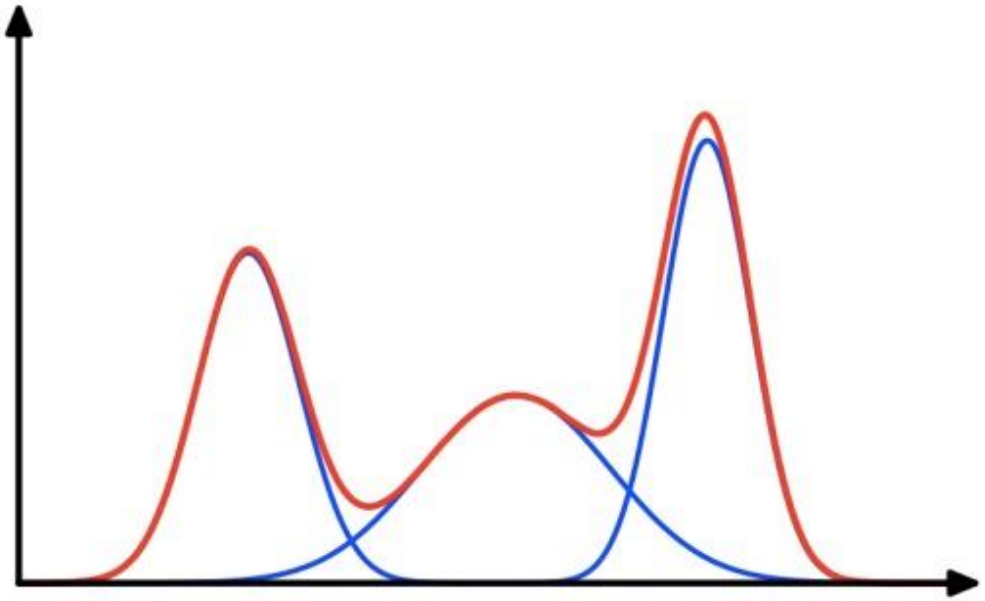


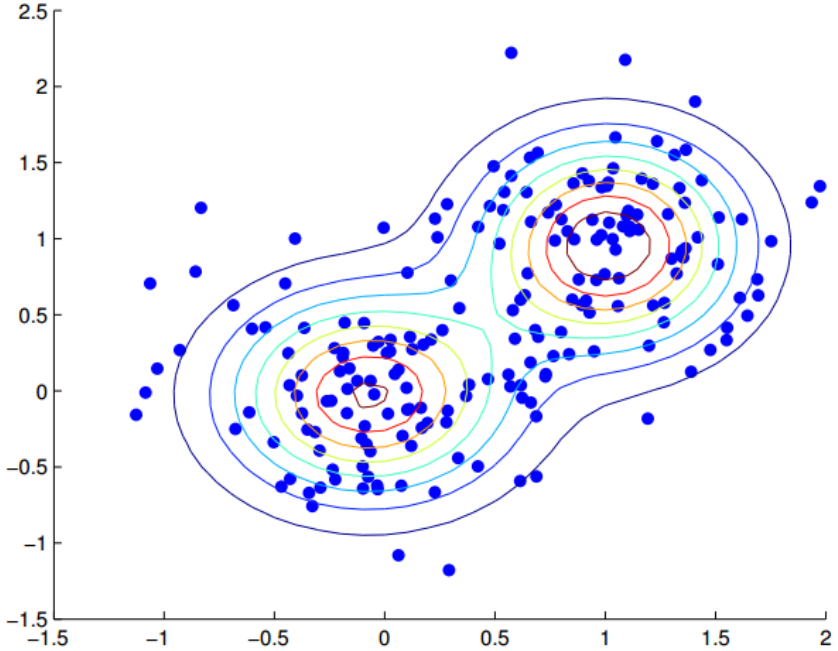


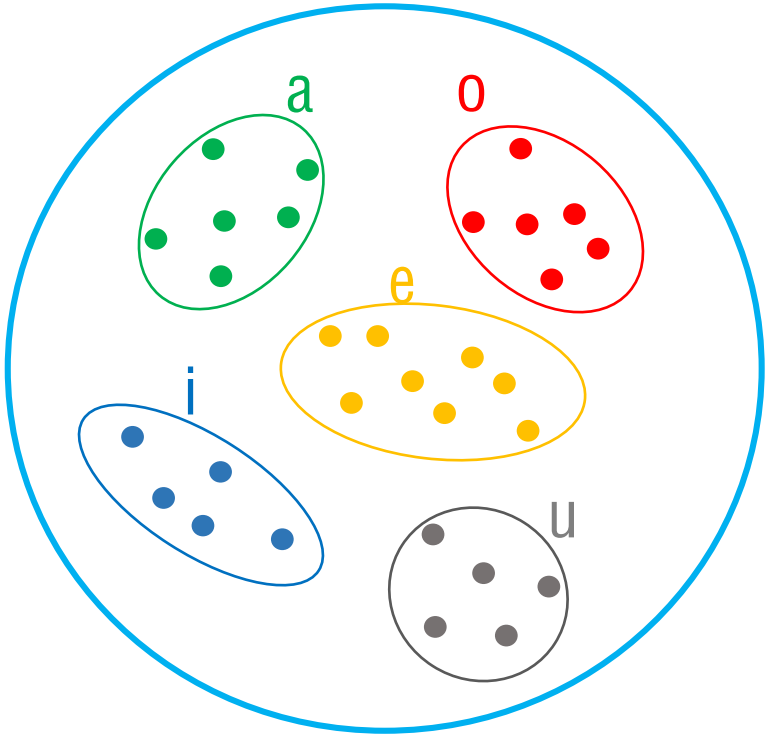
高斯模型就是用高斯概率密度函数（正态分布曲线）精确地量化事物，将一个事物分解为若干的基于高斯概率密度函数（正态分布曲线）形成的模型。对图像背景建立高斯模型的原理及过程：图像灰度直方图反映的是图像中某个灰度值出现的频次，也可以以为是图像灰度概率密度的估计。

高斯混合模型(GMM)

混合模型是一个可以用来表示在总体分布（distribution）中含有K个子分布的概率模型，换句话说，混合模型表示了观测数据在总体中的概率分布，它是一个由K个子分布组成的混合分布。混合模型不要求观测数据提供关于子分布的信息，来计算观测数据在总体分布中的概率。



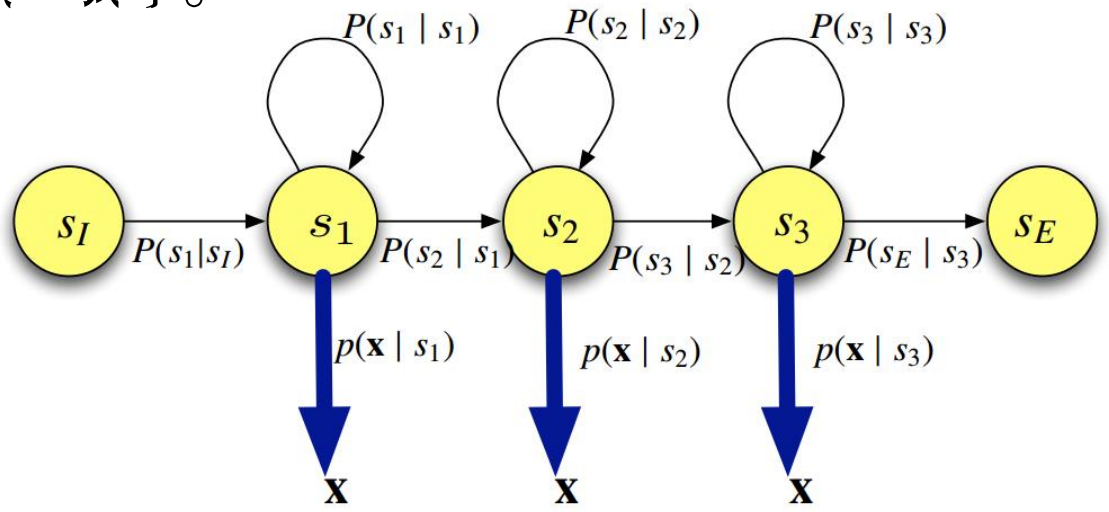




隐马尔可夫模型(HMM)

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。我们知道，对单个文字而言，虽然在不同语境下有轻微变音，但大致发音是有统计规律的。另一方面，当我们说出一句话时，文字与文字之间也是有一些转移规律的。比如，当我们说出“比”这个字时，下一个大概率的字一般是“如”“较”等。虽然文字千千万，但文字与文字之间的转移却是有章可循的。有了文字的发音特征，以及文字与文字之间的转移规律，那么从一段音频中推测出对应的文字也就可以一试了。



隐马尔可夫模型(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。

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。

隐马尔可夫模型(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：

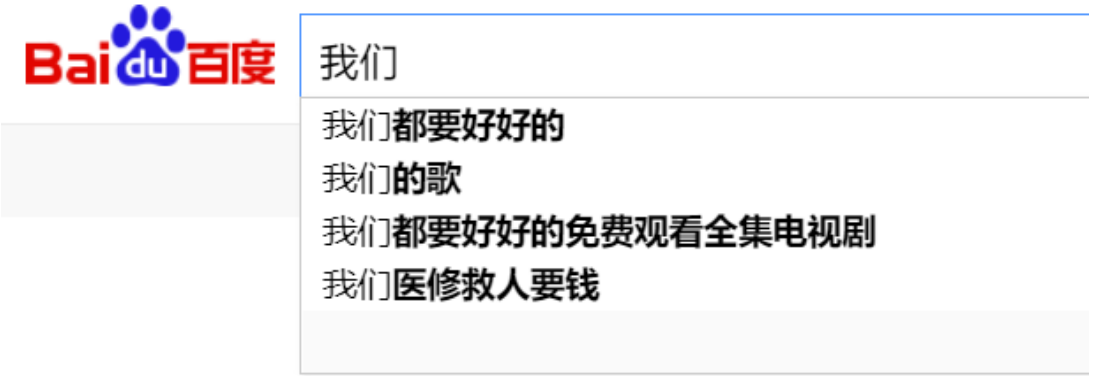
序列标注问题不仅和单个词相关，而且和观察序列的长度，单词的上下文，等等相关。

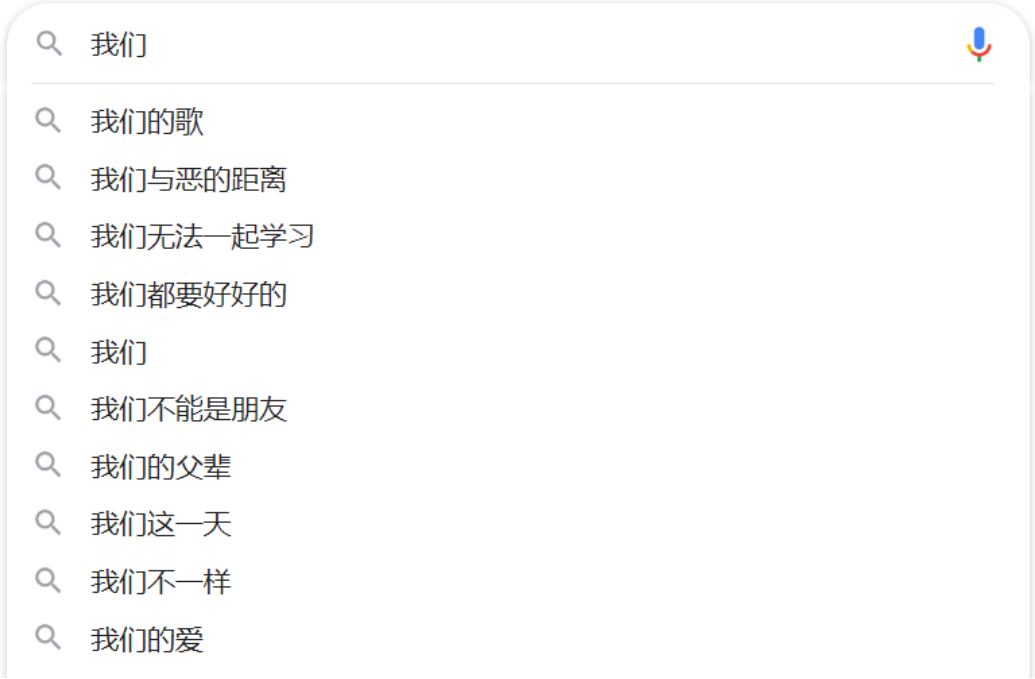
2、目标函数和预测目标函数不匹配：

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。

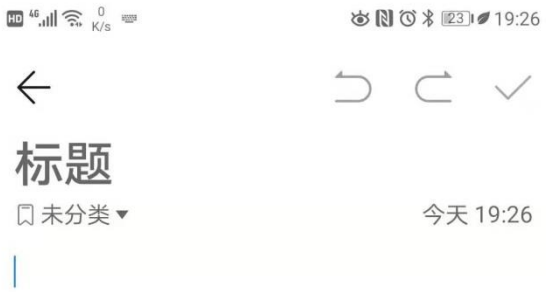
语言模型n-gram





简而言之，语言模型用来“猜”出哪些句子的可能性更大

语言模型n-gram



手机九宫格拼音输入法

按了2、3、4、5、6，出现很多备选，排在前面的可能性更大。

排名依据

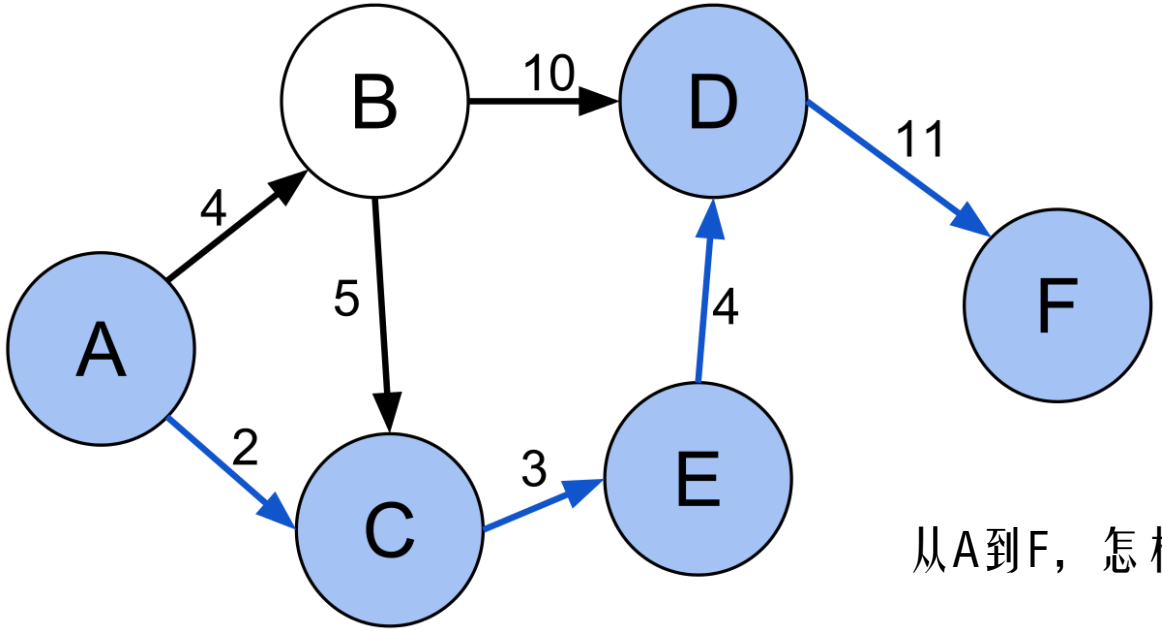
- 该词是否更加“正常”、更加常用
- 用户是否常用该词（可能别人不常用，但系统能够记忆用户的习惯）

语言模型n-gram

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

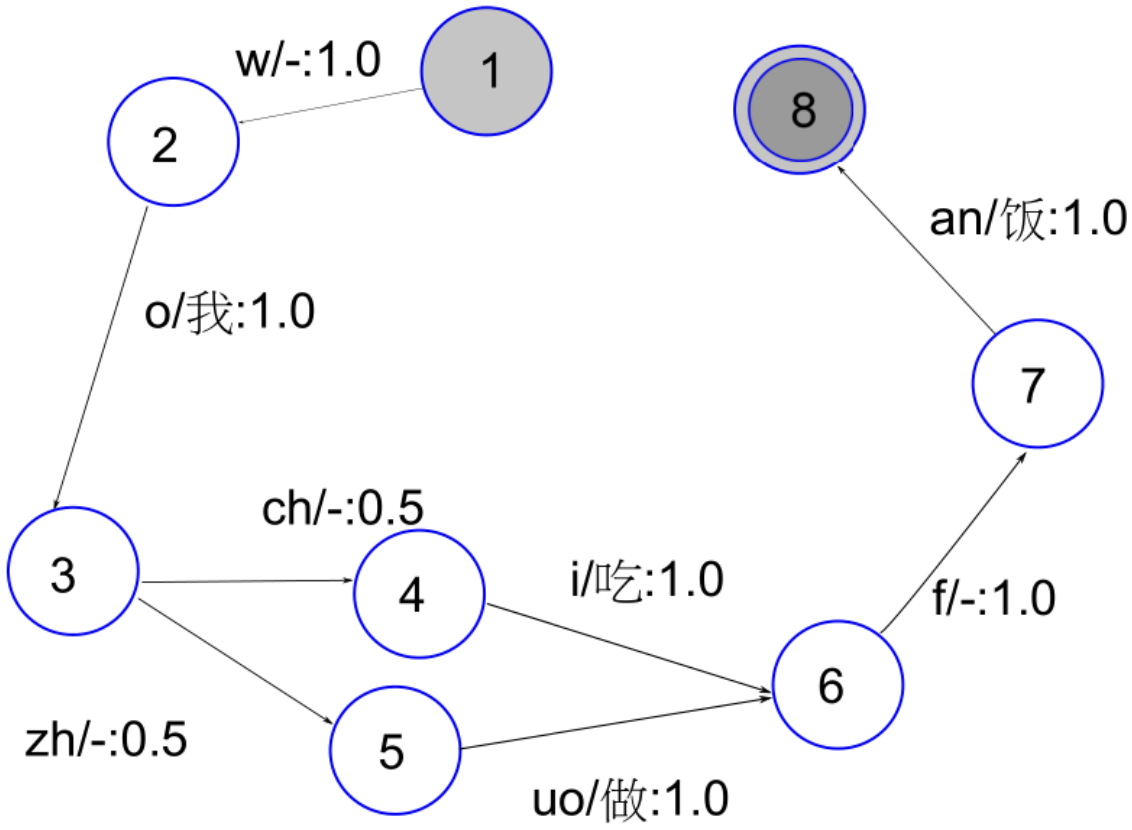
解码过程



从A到F，怎样走得分最高。

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。

解码过程

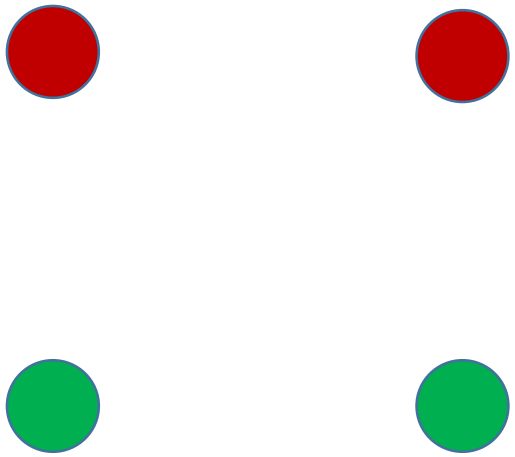


目录

- 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

深度学习

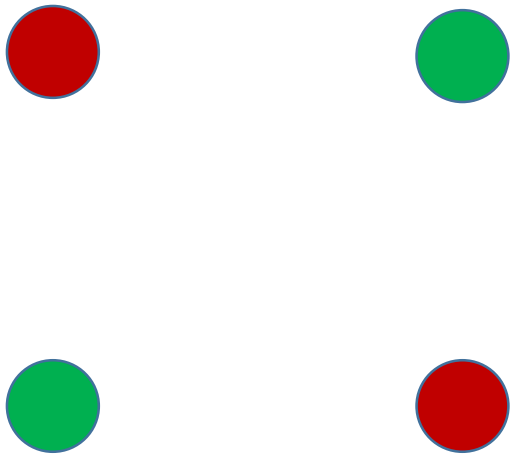
一个简单的分类问题(线性可分)

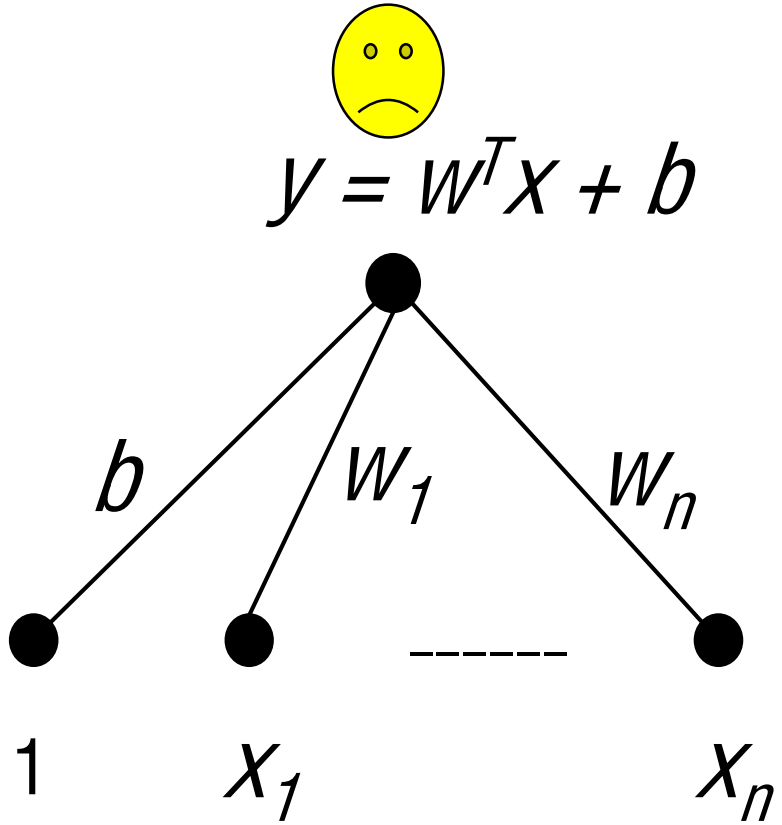


$$y = w^T x + b$$

深度学习

另一个分类问题(异或)

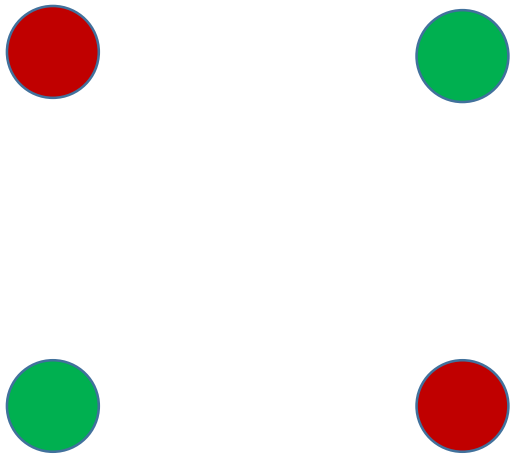


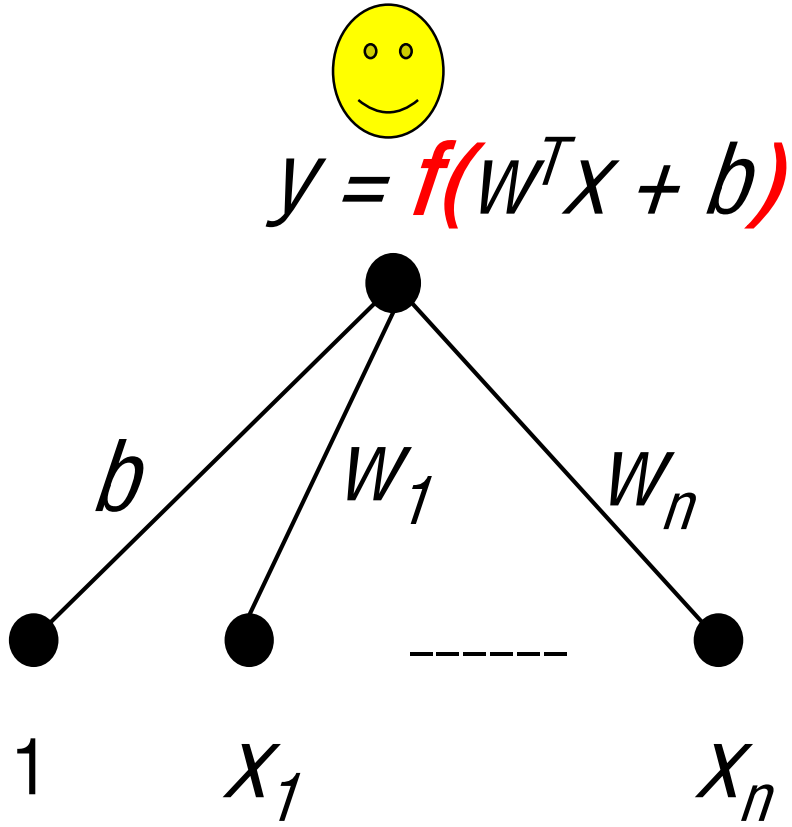


深度学习

神经网络的基本结构

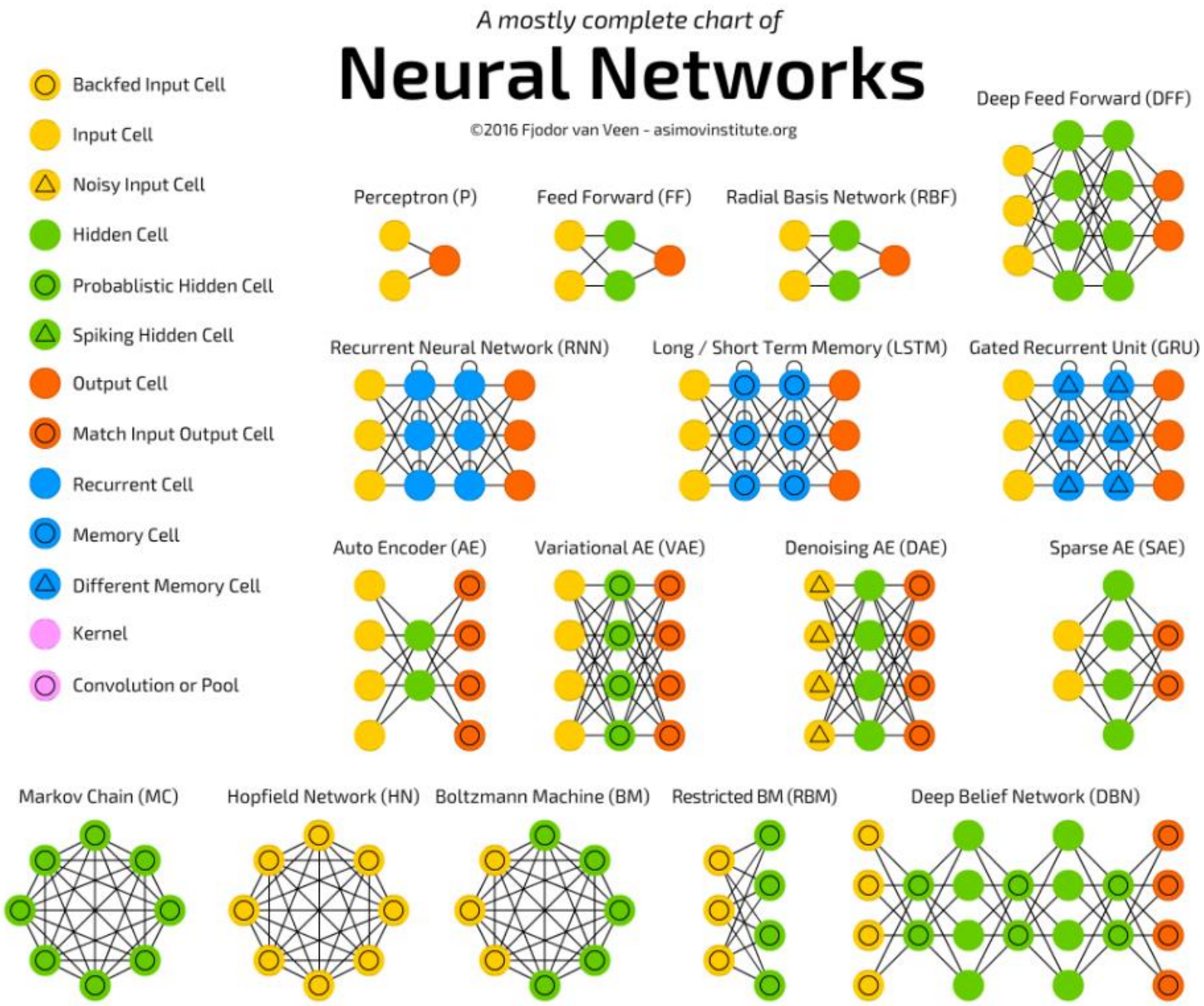
备注：
既然是非线性分类问题，将模型引入非线性函数即可





深度学习

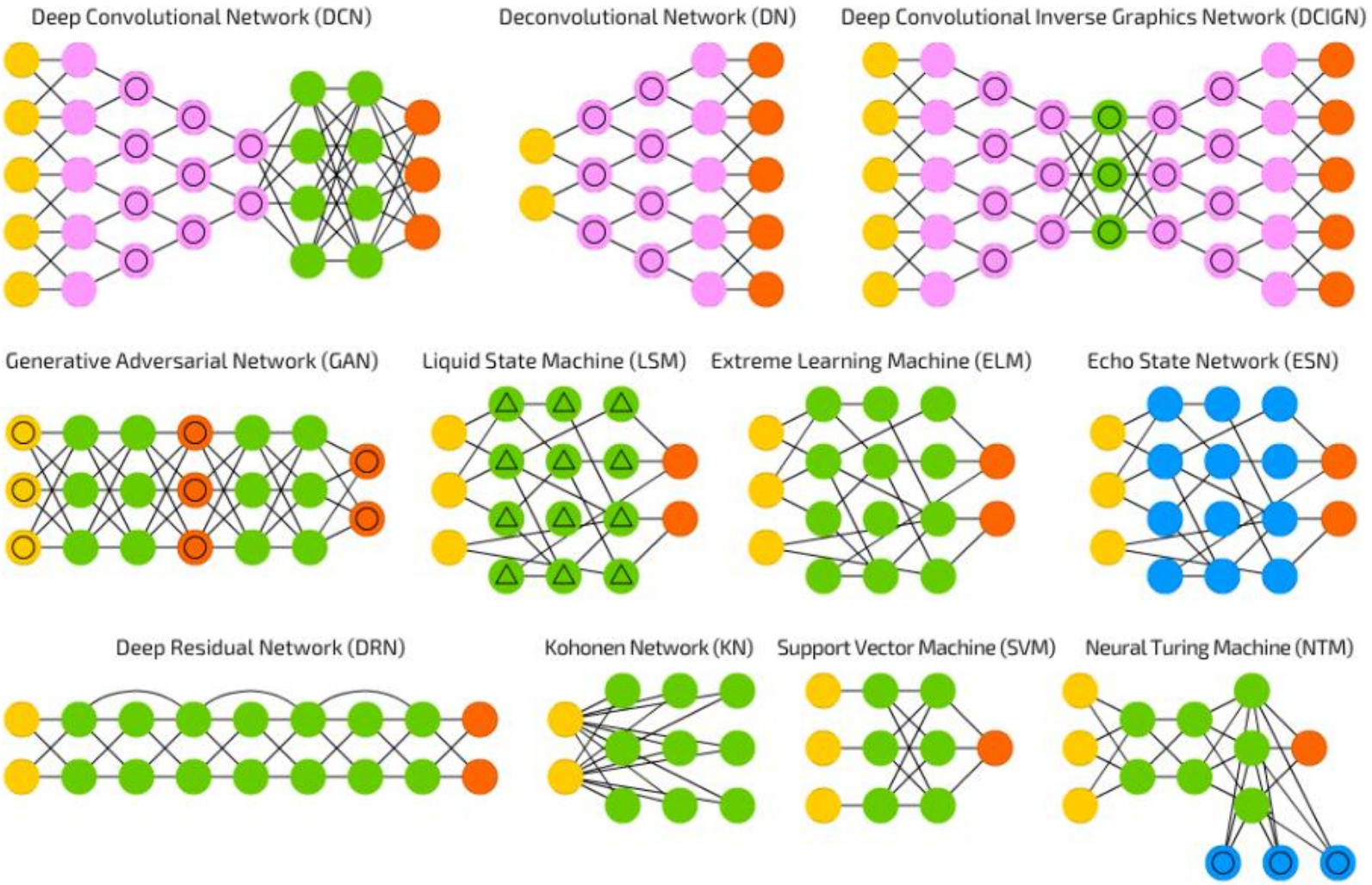
网络结构


A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)

深度学习

网络结构

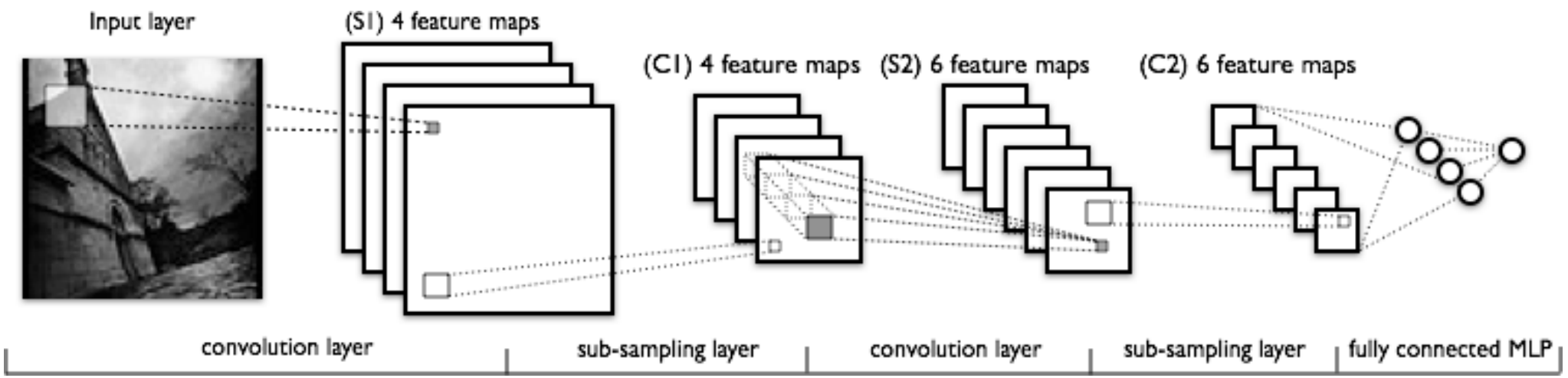
备注：
神经网络的灵活性体现在不同的网络结构



深度学习

卷积神经网络(CNN)

备注：
最常用的网络之卷积神经网络，它的设计初衷是模拟人的视网膜细胞，每个细胞只对部分区域负责，各个细胞各有分工而又相互联系

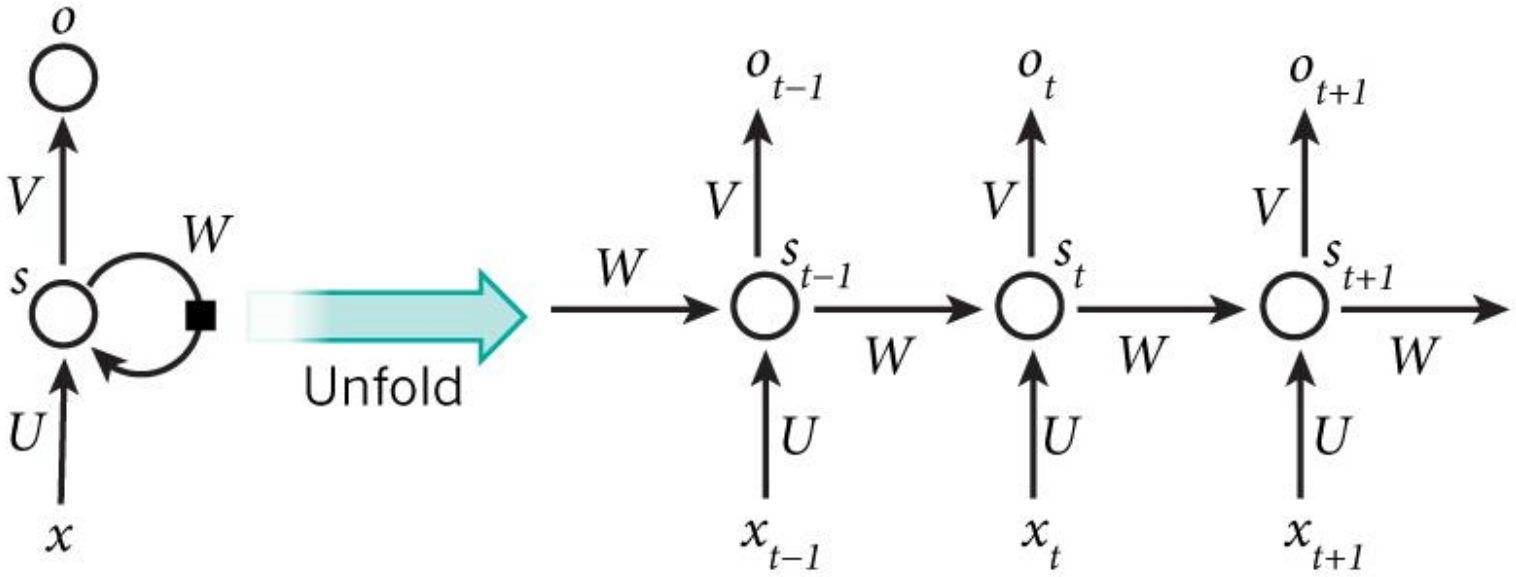


深度学习

循环神经网络(RNN)

备注：
最常用的网络之循环神经网络，
铭记历史能够相助助于当下

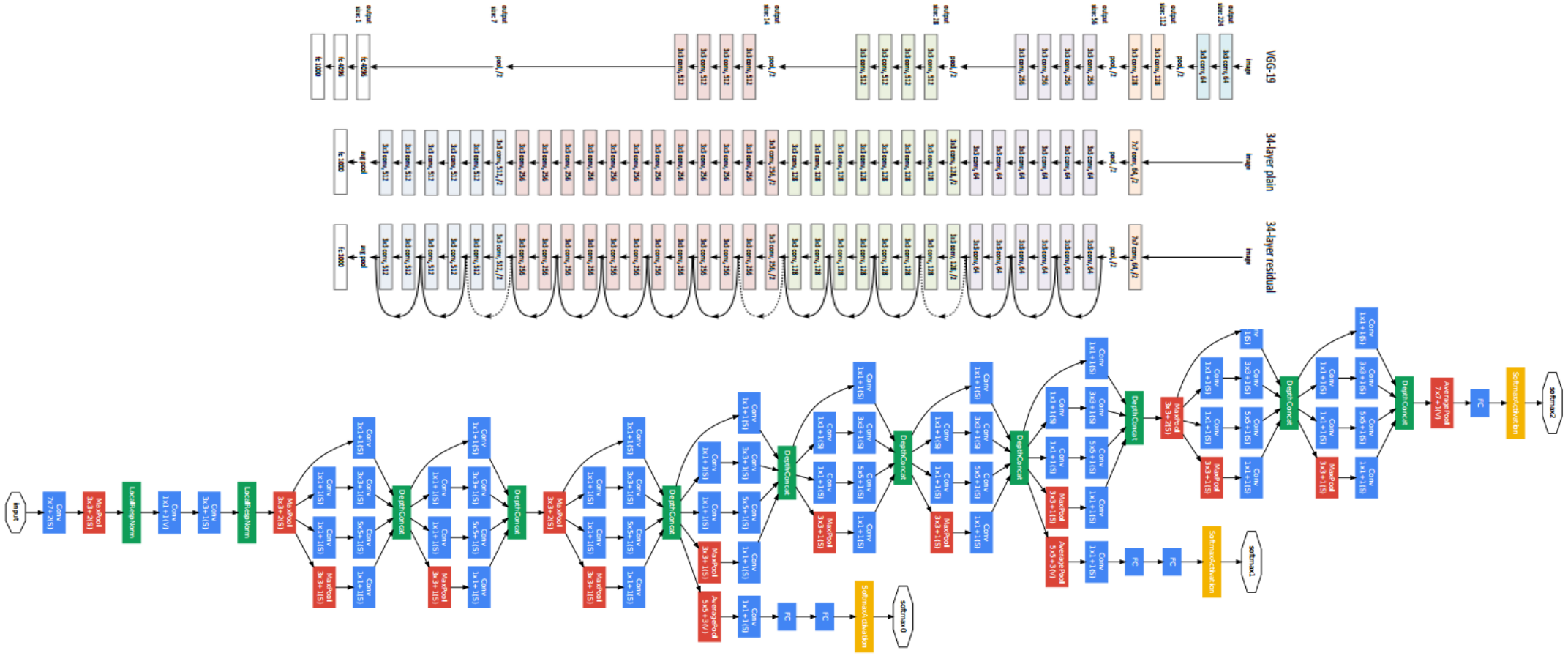
在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。RNNs之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。



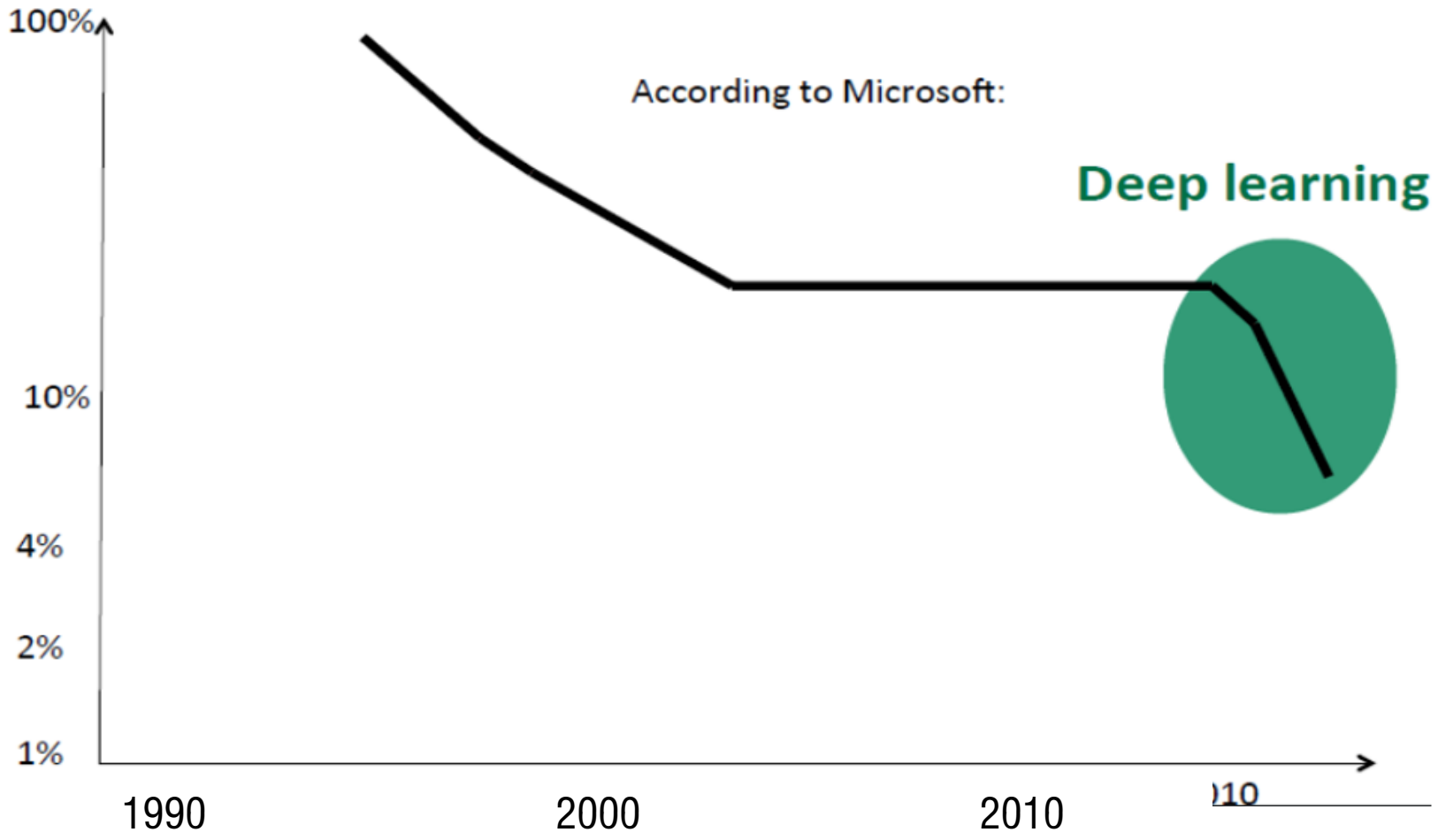
深度学习

深度神经网络(来自微软、谷歌)

备注：
结合上面的几项可选性，可以构造很复杂的网络，下图两个网络是横着放的

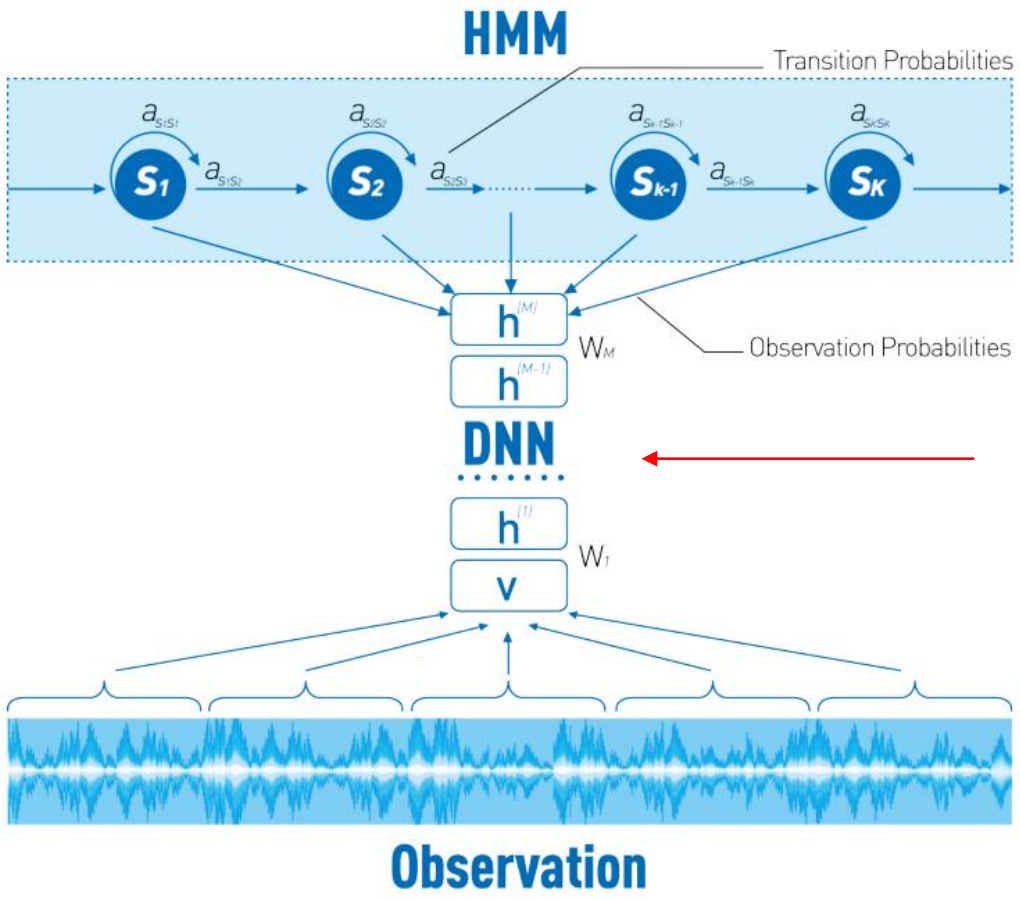

VGG-19
34-layer plain
34-layer residual
input
Conv 7x7+ 2(S)
MaxPool 3x3+ 2(S)
LocalRespNorm
Conv 1x1+ 1(V)
Conv 3x3+ 1(S)
LocalRespNorm
MaxPool 3x3+ 2(S)
Conv 1x1+ 1(S)
Conv 3x3+ 1(S)
Conv 5x5+ 1(S)
MaxPool 3x3+ 1(S)
DepthConcat
Conv 1x1+ 1(S)
MaxPool 3x3+ 1(S)
DepthConcat
MaxPool 3x3+ 2(S)
Conv 1x1+ 1(S)
Conv 3x3+ 1(S)
Conv 5x5+ 1(S)
MaxPool 3x3+ 1(S)
DepthConcat

语音识别遇见深度学习



纵轴是错误率，随着技术的发展，语音识别系统错误率越来越低，甚至低于人类。

语音识别遇见深度学习

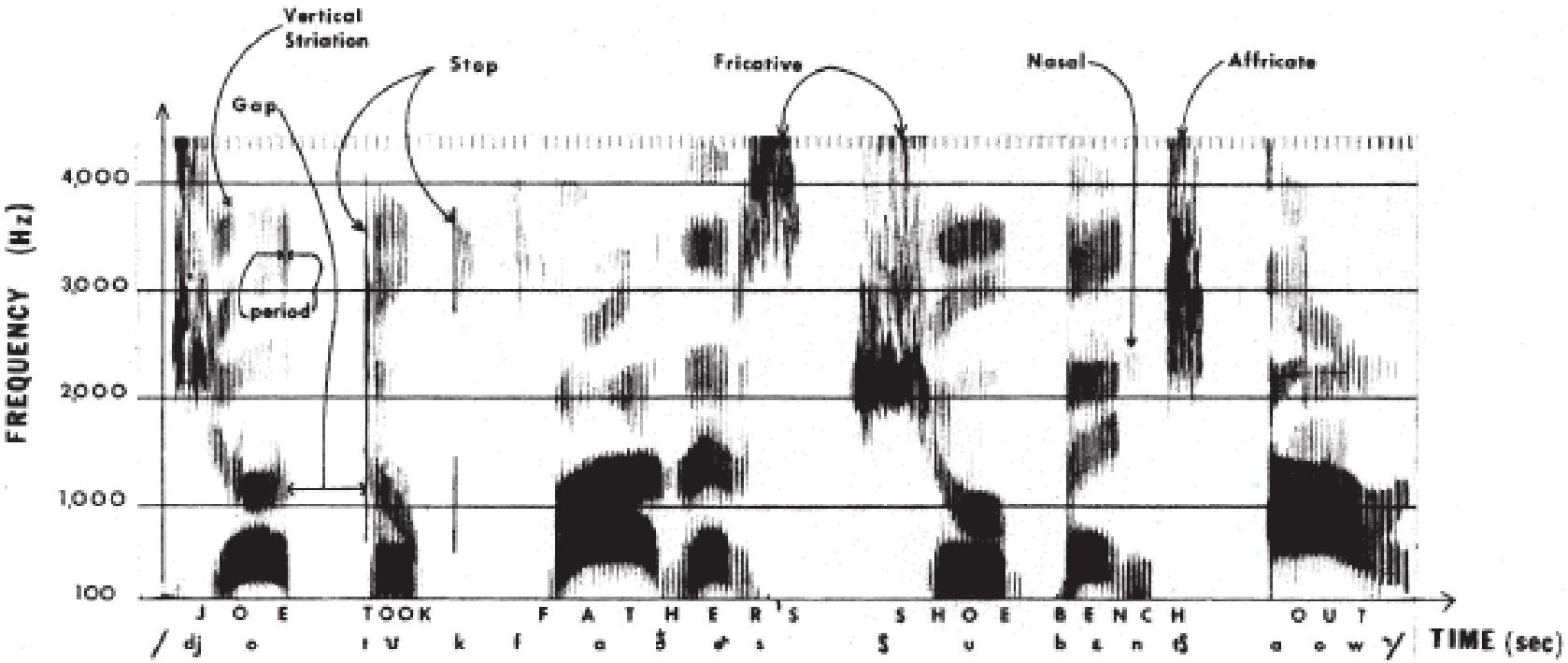


DNN替换了前文的GMM

目录

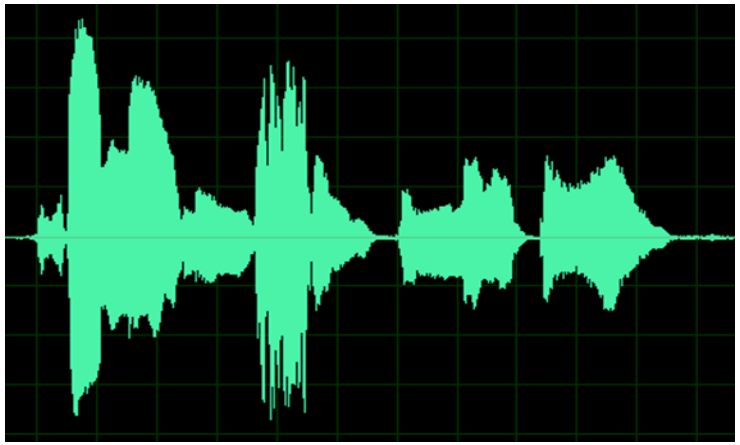
- 语音的产生与感知
- 语音识别概述
- 基于GMM-HMM的语音应用
- 基于深度学习的语音识别
- 说话人识别
- 语音合成
- 语音的应用

声纹

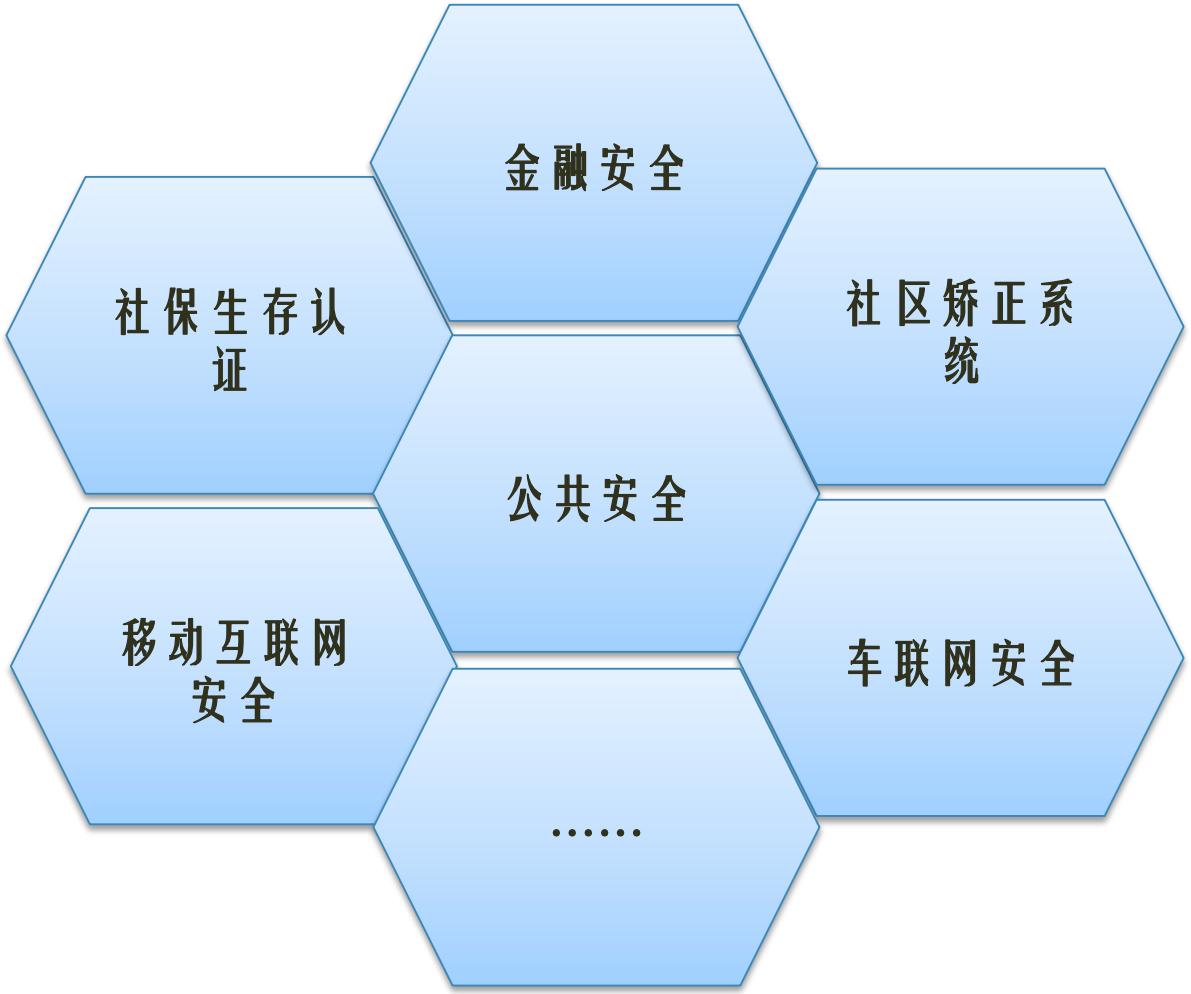

Vertical Striation
Stop
Fricative
Nasal
Affricate
Gap
period
FREQUENCY (Hz)
4,000
3,000
2,000
1,000
100
J O E
TOOK
F A T H E R'S
S H O E
B E N C H
O U T
TIME (sec)

声纹识别

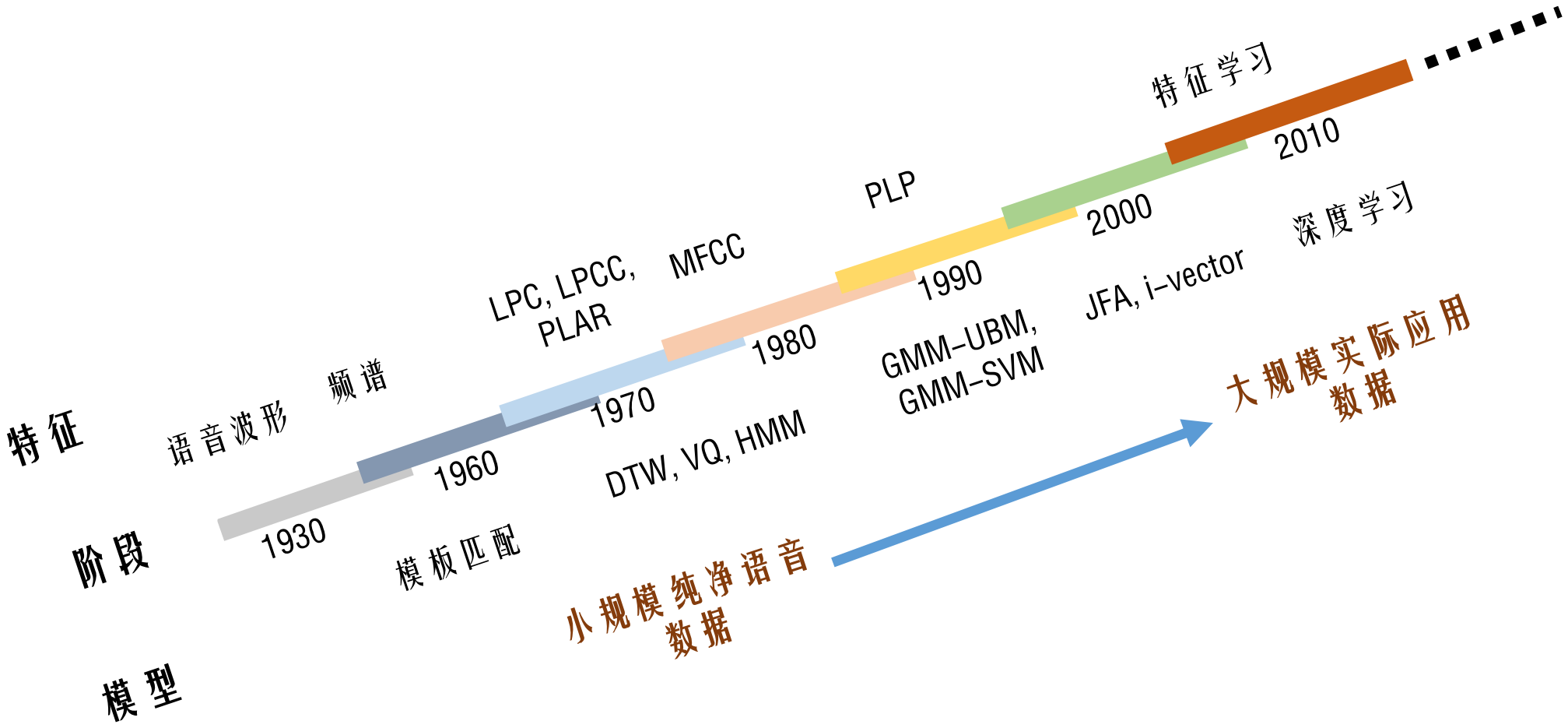
- **声纹（Voiceprint）：**语音中所蕴含的唯一表征说话人身份的语音特征参数及基于这些特征参数所建立的语音模型。
- **声纹识别（Voiceprint Recognition，或**说话人识别，Speaker Recognition)：基于语音中包含的说话人特有的个性信息自动鉴别语音对应的说话人身份。



声纹识别应用领域

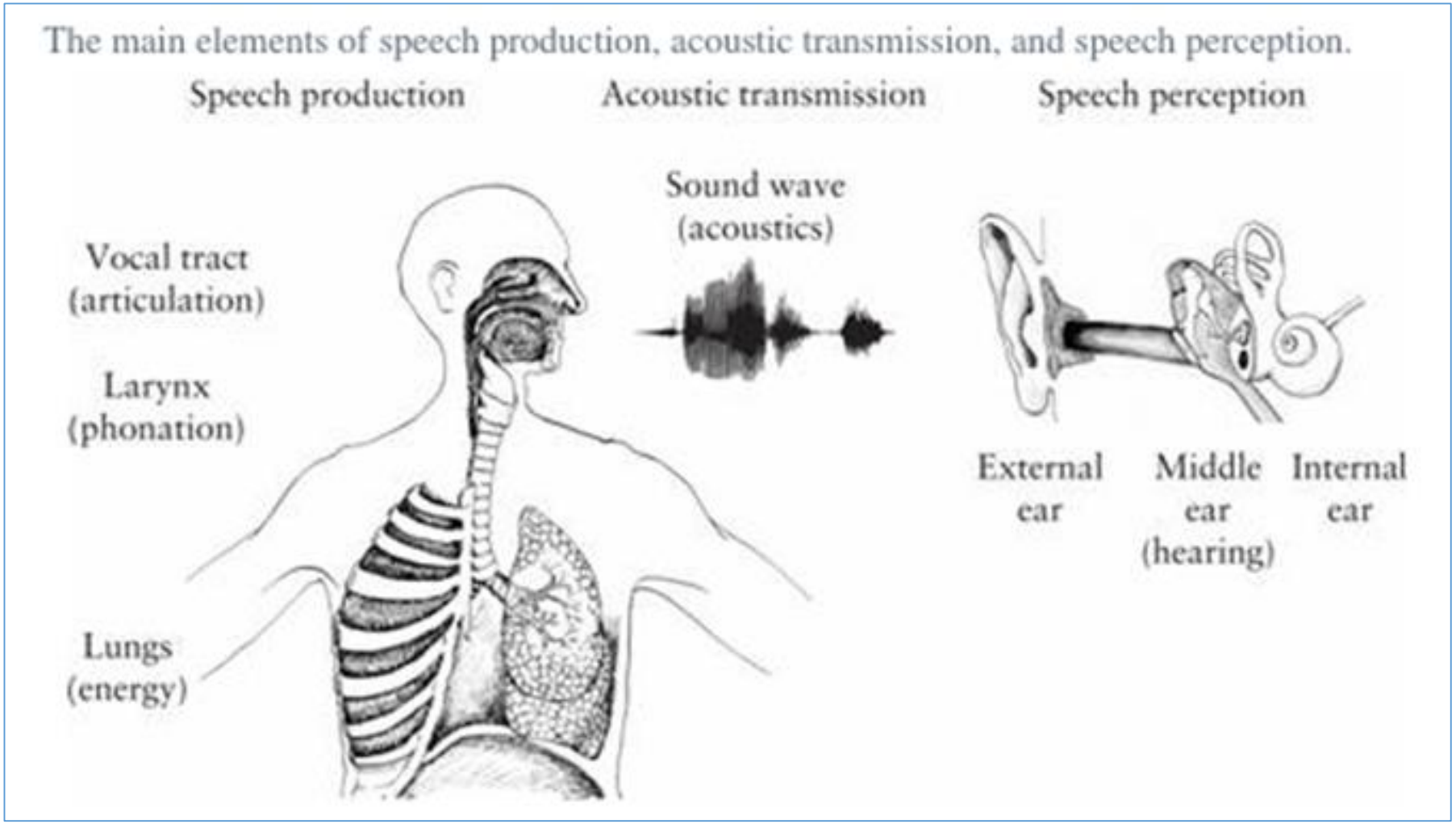


声纹识别技术发展

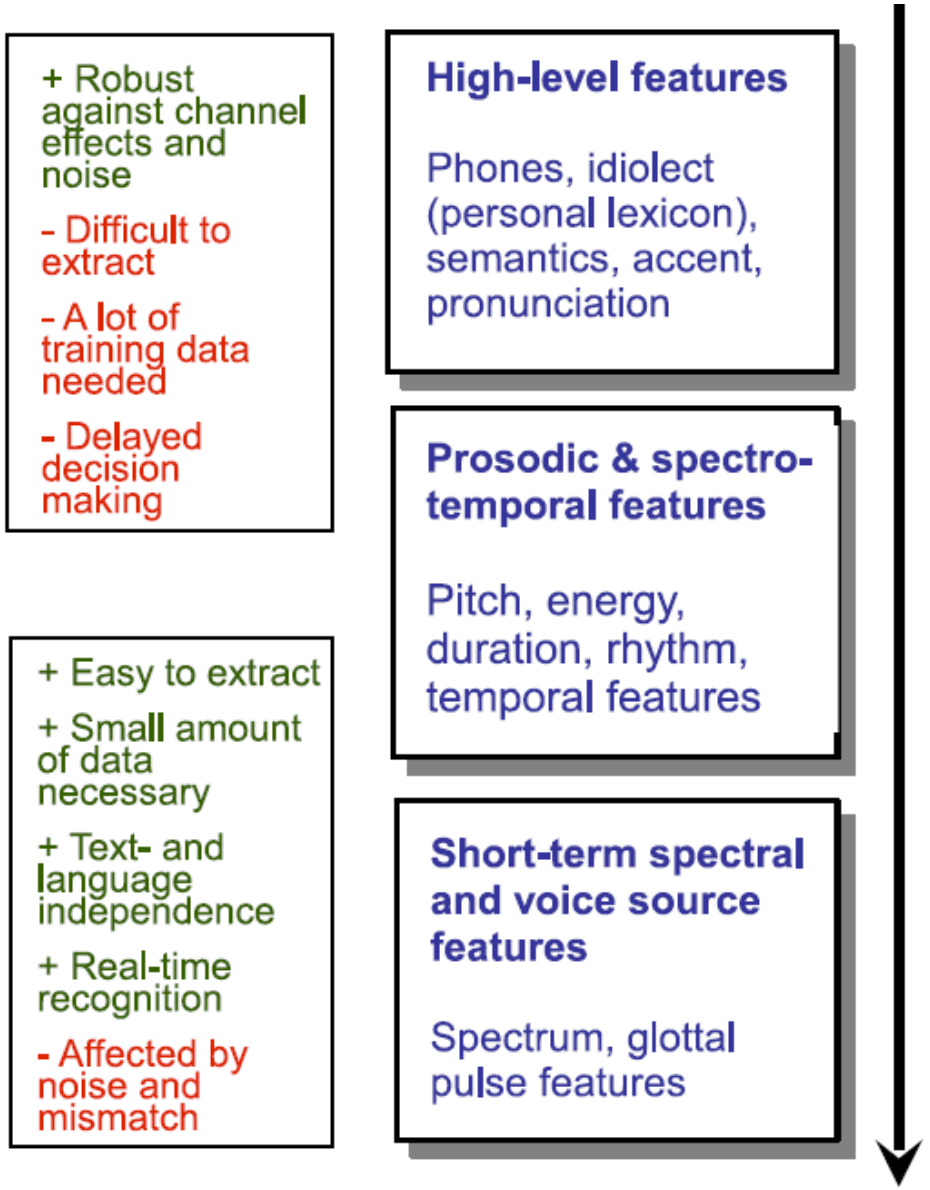

特征
阶段
模型
语音波形
频谱
LPC, LPCC, PLAR
MFCC
PLP
特征学习
1930
1960
1970
1980
1990
2000
2010
模板匹配
DTW, VQ, HMM
GMM-UBM, GMM-SVM
JFA, i-vector
深度学习
小规模纯净语音数据
大规模实际应用数据

研究阶段（一）特征提取

- 发现能代表说话人信息的典型特征
- 关注语音产生机理及感知机理



研究阶段（一）特征提取



Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence

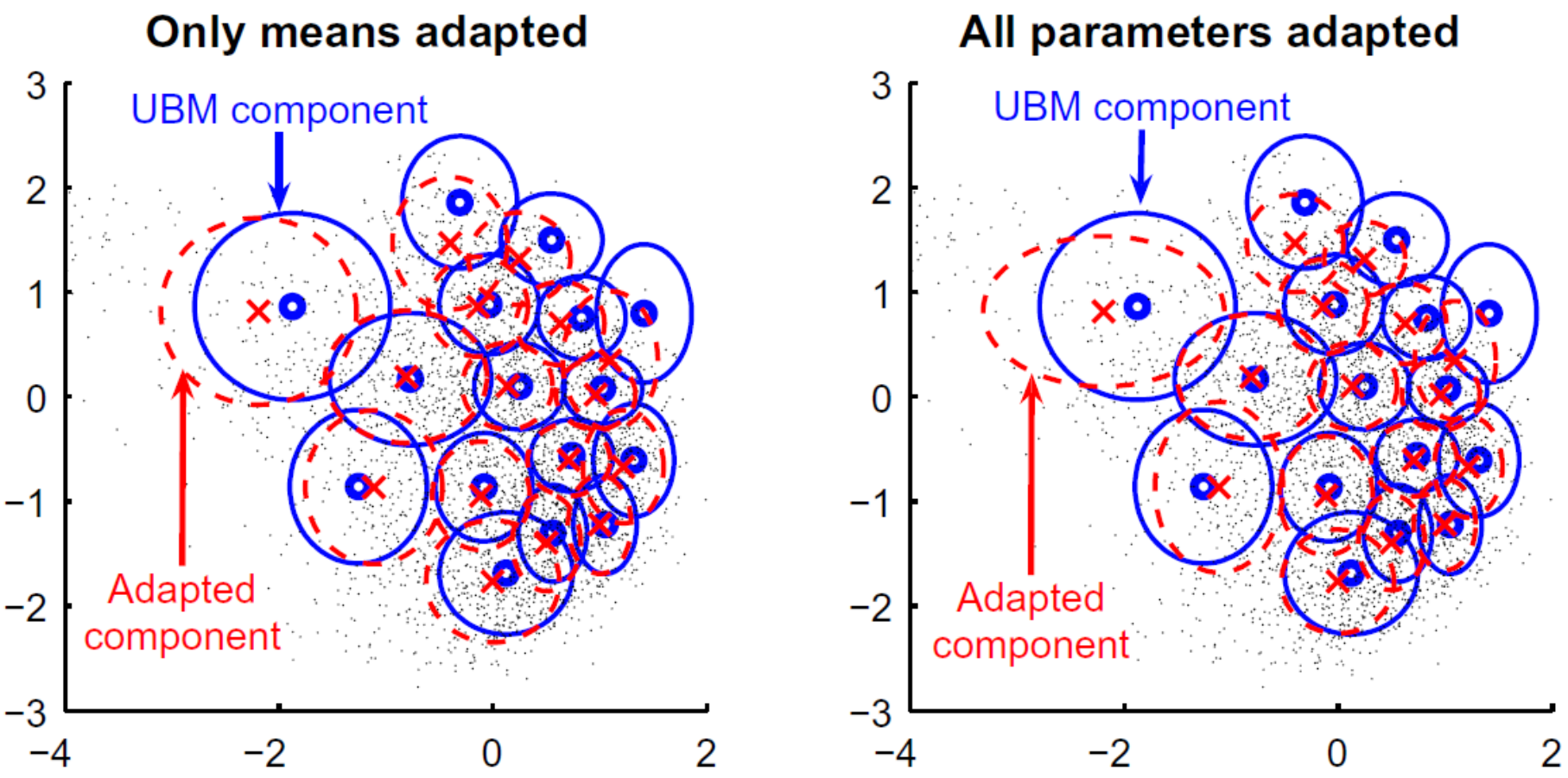
后天学习的

Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract

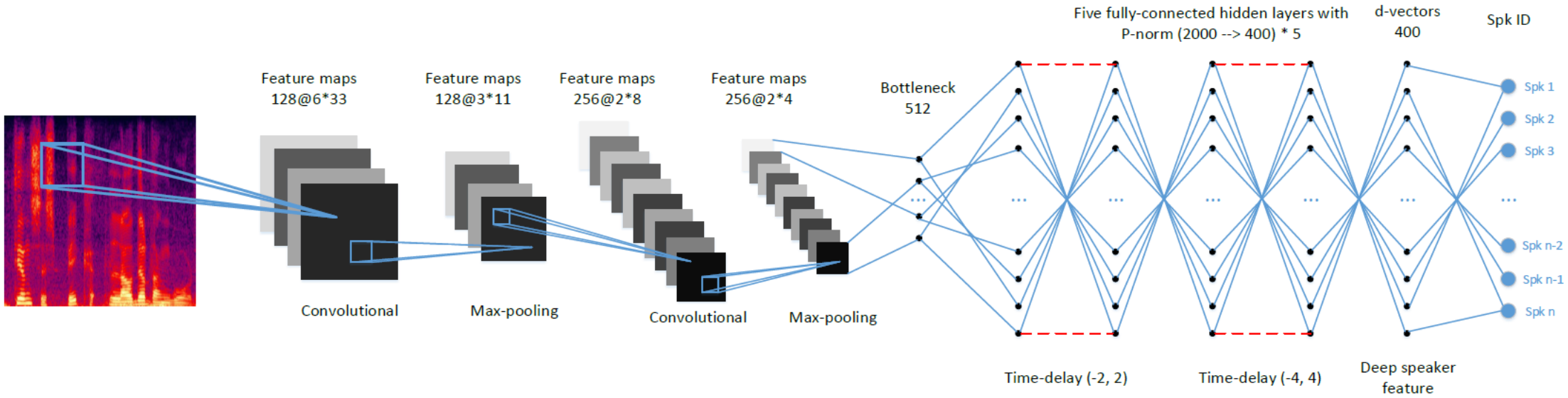
生来就有的

研究阶段（二）统计模型

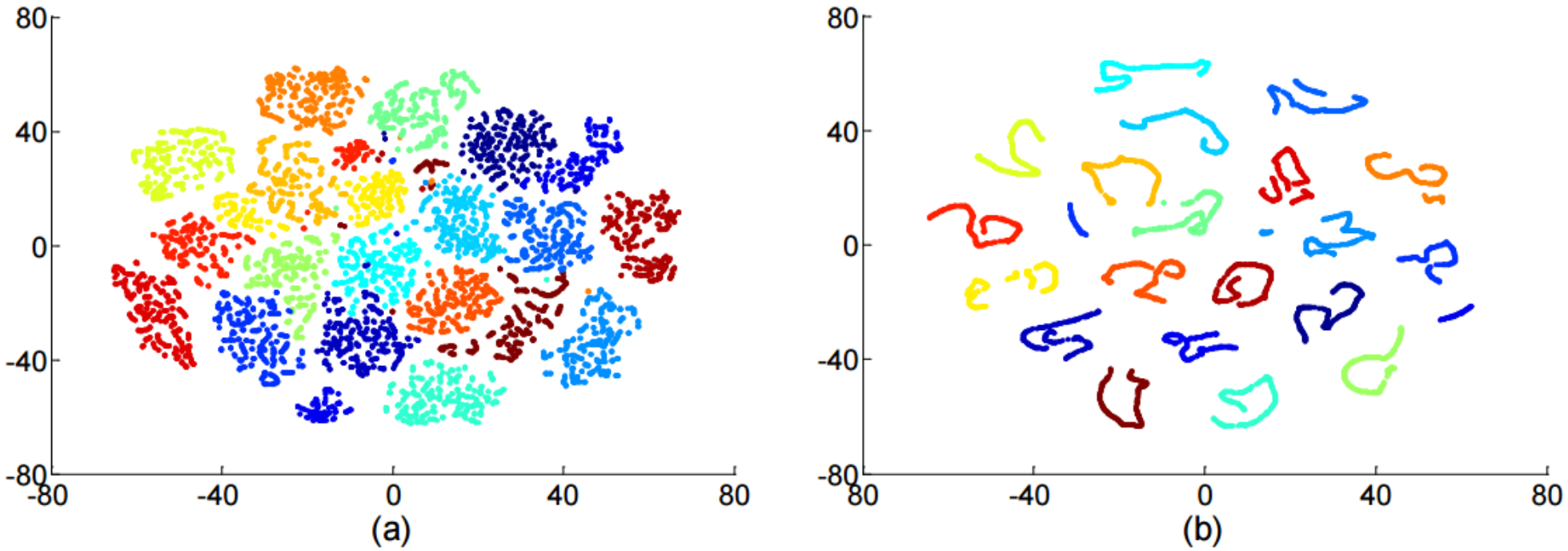
- 高斯模型-通用背景模型（GMM-UBM）



研究阶段（三）深度学习


Feature maps
128@6*33
Feature maps
128@3*11
Feature maps
256@2*8
Feature maps
256@2*4
Bottleneck
512
Five fully-connected hidden layers with
P-norm (2000 --> 400) * 5
d-vectors
400
Spk ID
Spk 1
Spk 2
Spk 3
Spk n-2
Spk n-1
Spk n
Convolutional
Max-pooling
Convolutional
Max-pooling
Time-delay (-2, 2)
Time-delay (-4, 4)
Deep speaker
feature

研究阶段（三）深度学习



在极短语音内即可学习到说话人的特征
（不同的颜色代表不同的说话人）

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(历史)

1）参数合成

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(历史)

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(历史)

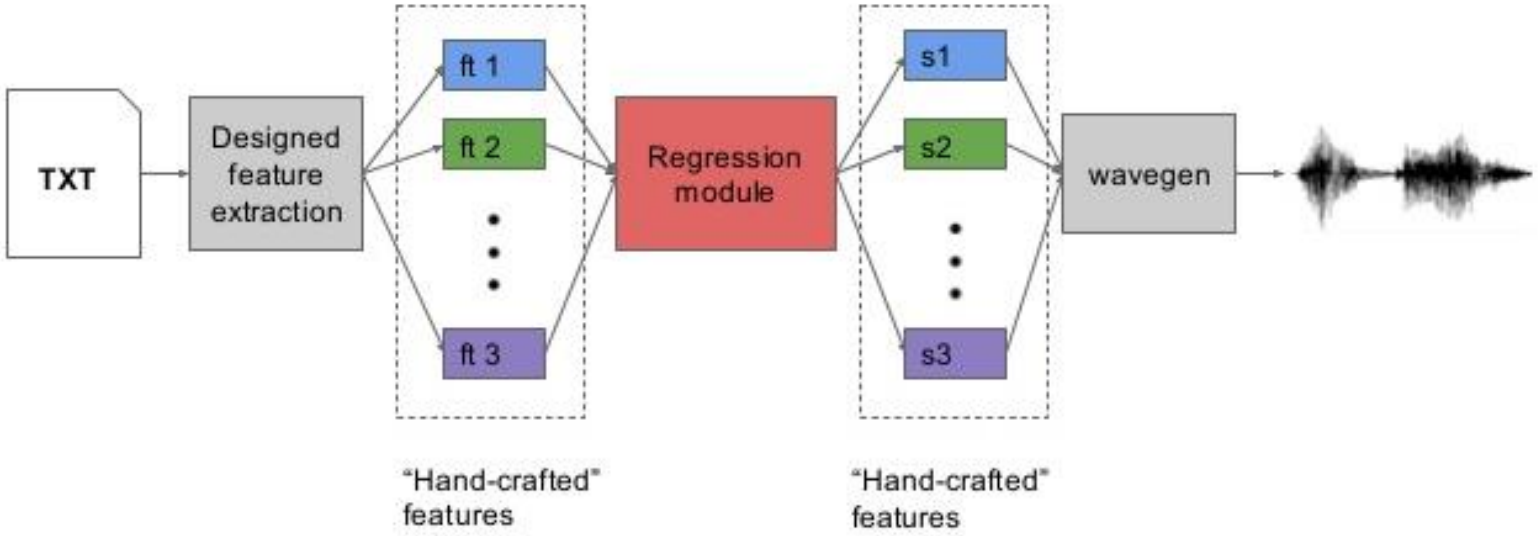
3）统计模型合成

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(历史)

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。



语音合成示例

机器合成古典音乐

机器合成人声

语音合成定制

科大讯飞 https://www.xfyun.cn/services/online_tts

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

语音的应用

智能家居控制系统

个性化配音服务

物联网输入法

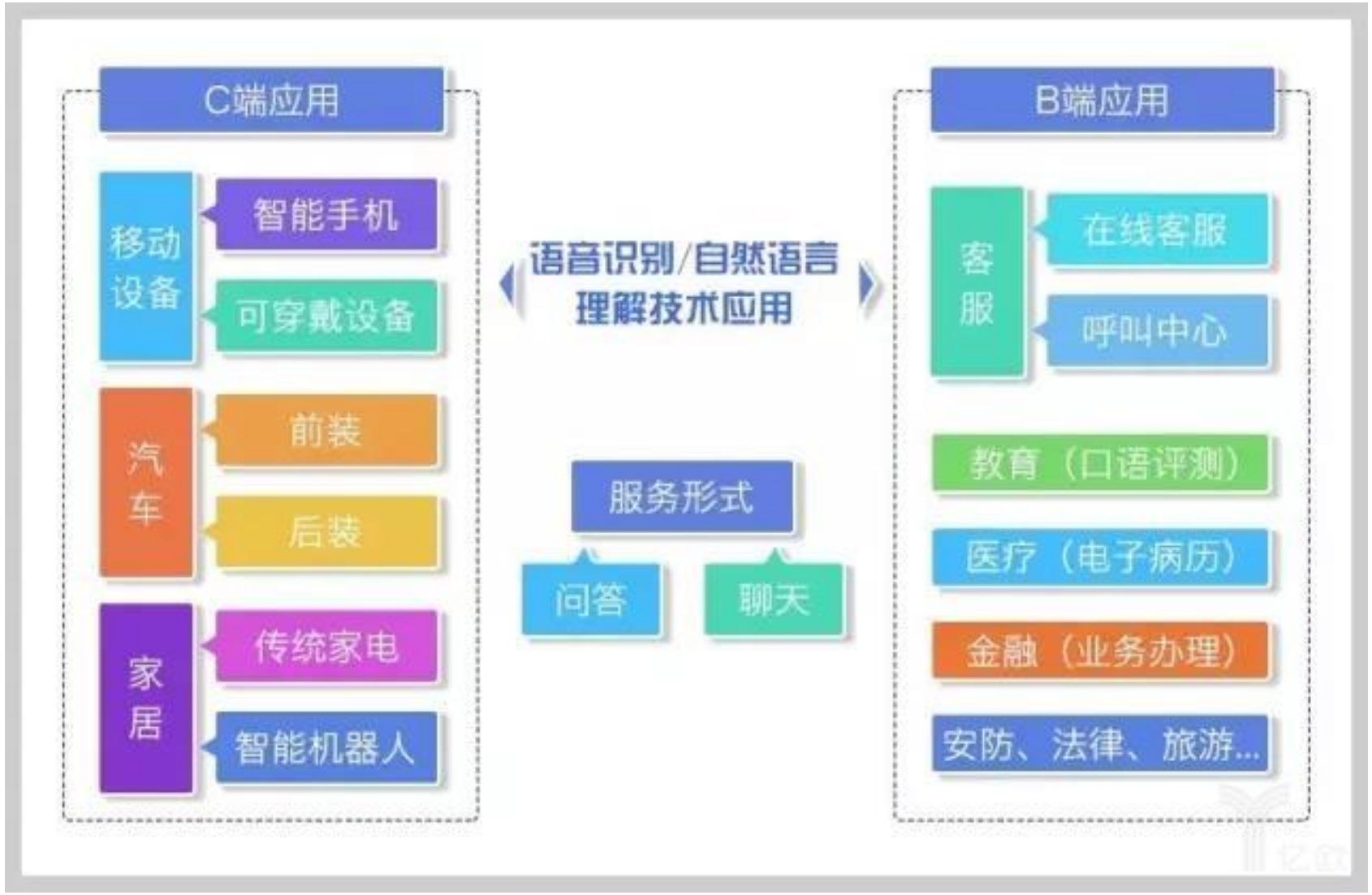
个性化语音合成

中英互译（其他语言）

智能客服机器人

汽车智能车载系统

语音的应用


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

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）。

他们竟然自己对话了…

The end