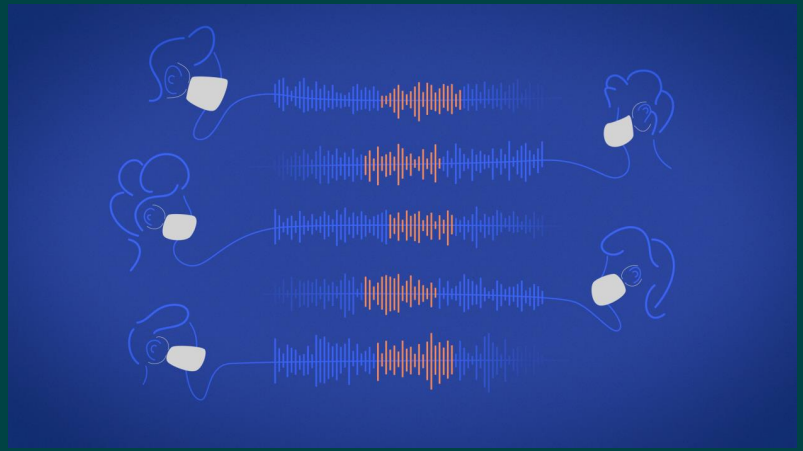
《人工智能导论》

主讲老师： 李蓝天

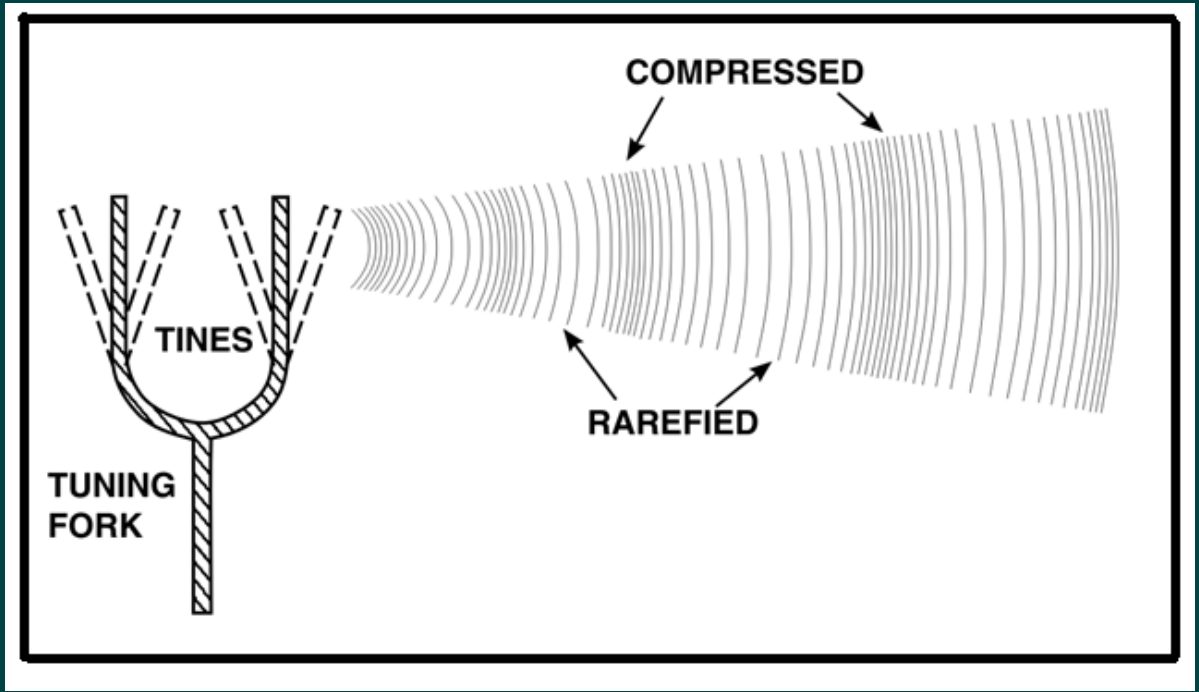
课程内容

1. 神奇的人工智能
2. 认识你的脸
3. 倾听你的声音
4. 模仿你的行为



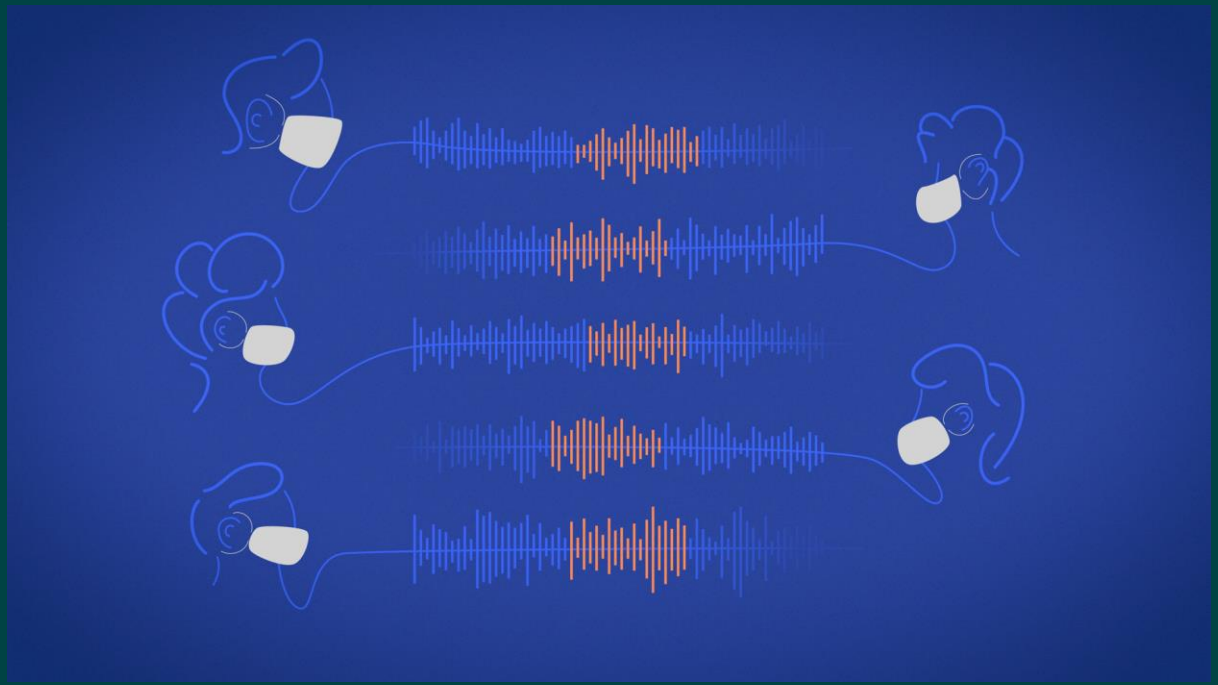
什么是声音?

声音是声源产生的振动，是空气中疏密相间的周期性变化



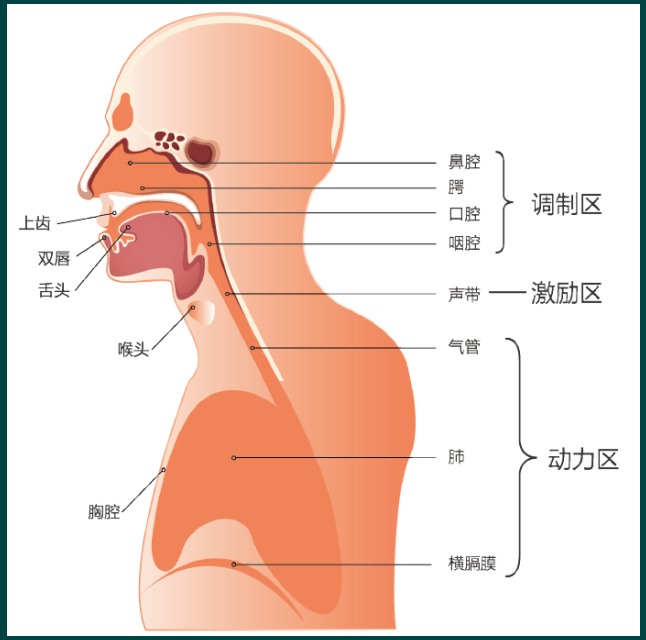
你的声音：语音

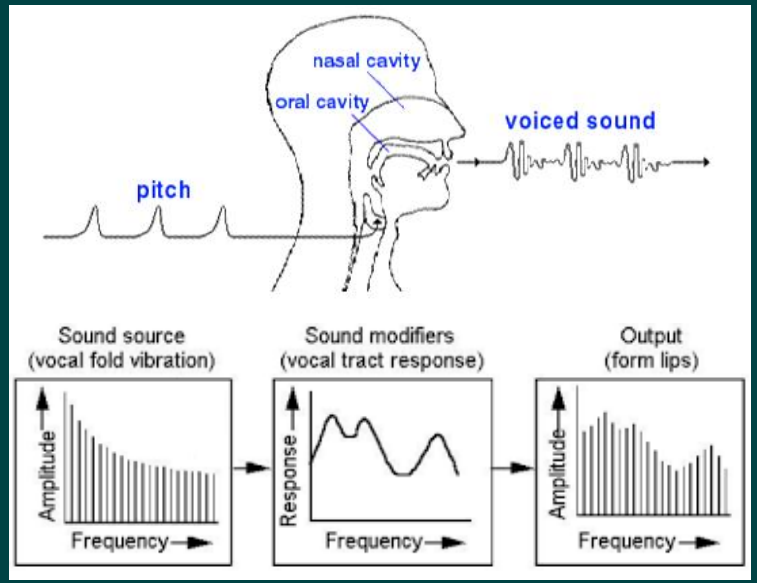
特指人类在交流中所发出的声音，便是语音。



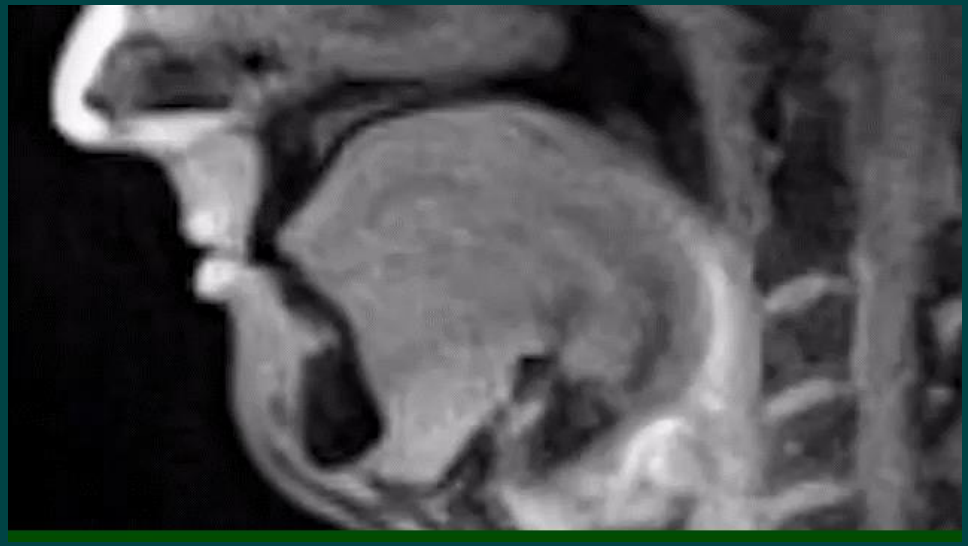
语音产生

激励-调制模型：由肺部产生气流，冲击声带产生振动（激励），引起气流疏密变化，经过口腔和鼻腔调制后，由口唇辐射到空气中，产生语音。



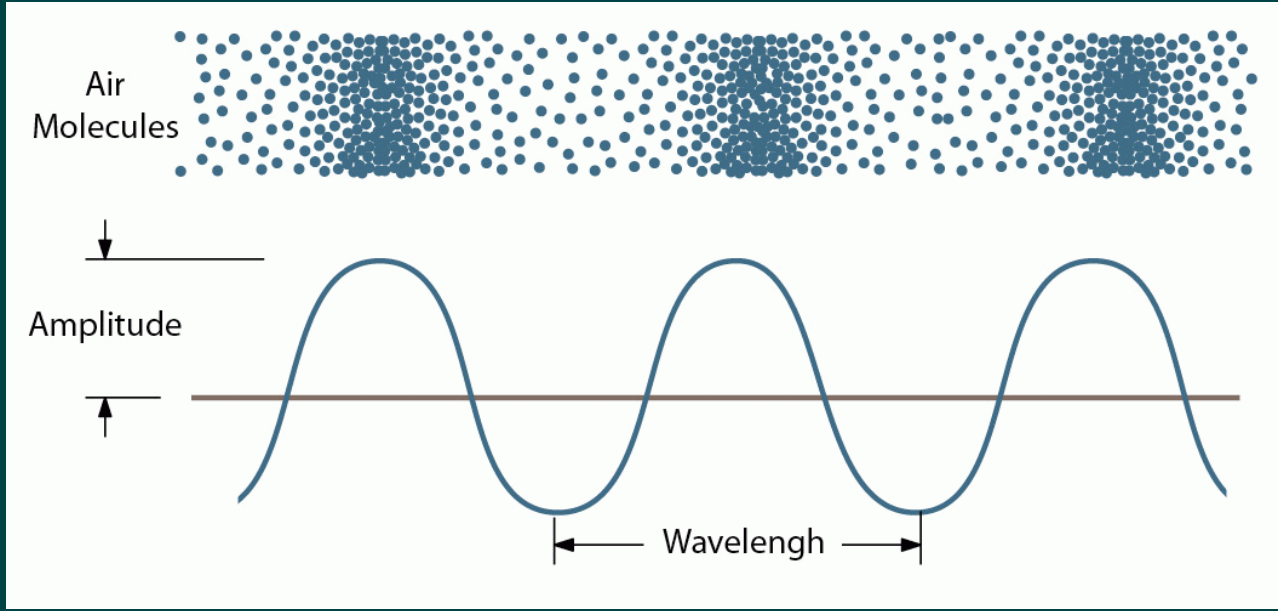


语音产生



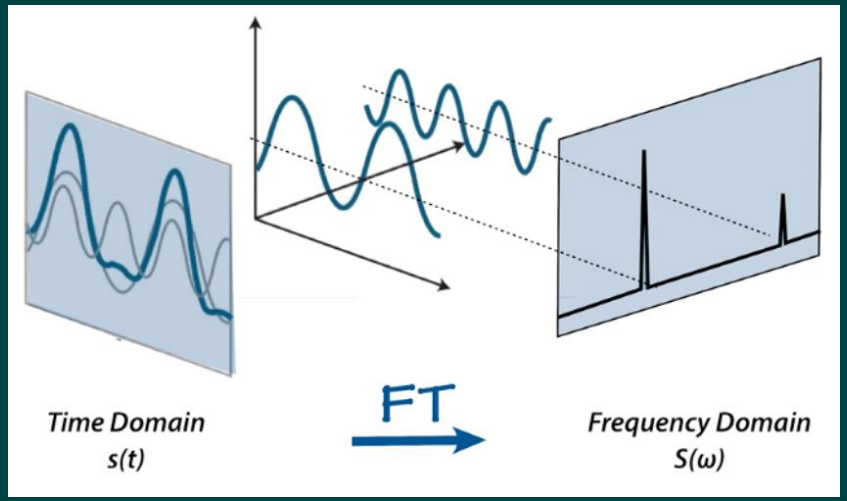
语音的模样：波形图

语音波形记录了空气密度的变化。某一时刻的空气密度代表了一个语音信号的采样点。若将密度值表示为时间的函数，则得到了语音波形图。

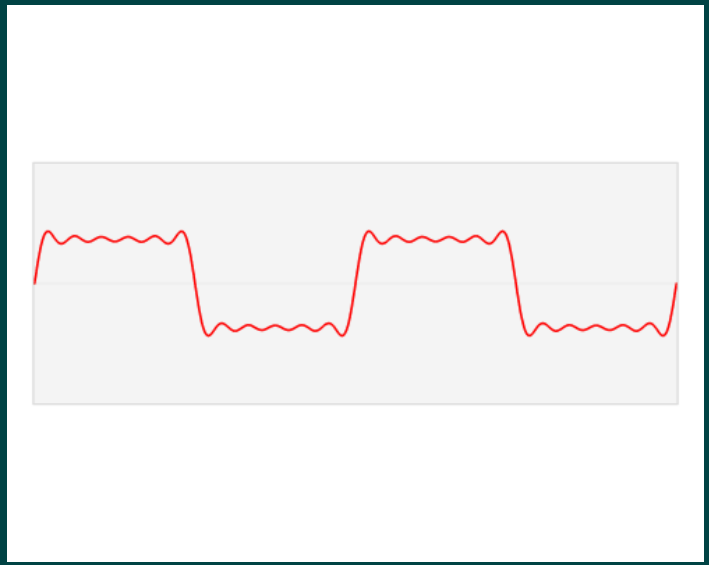


语音的模样：频谱图

语音信号具有短时平稳性。因此，可以分帧进行短时分析。每一帧在不同频率上的能量称为该帧的频谱。



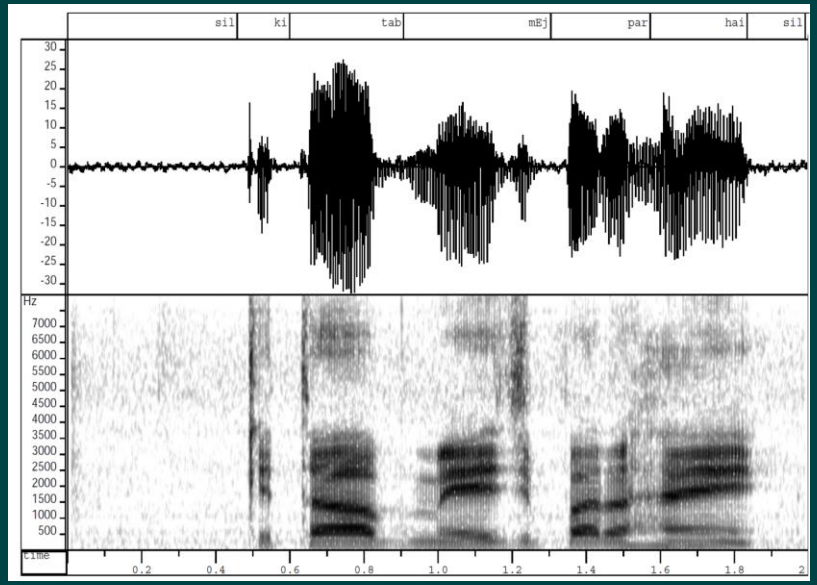
短时傅里叶变换

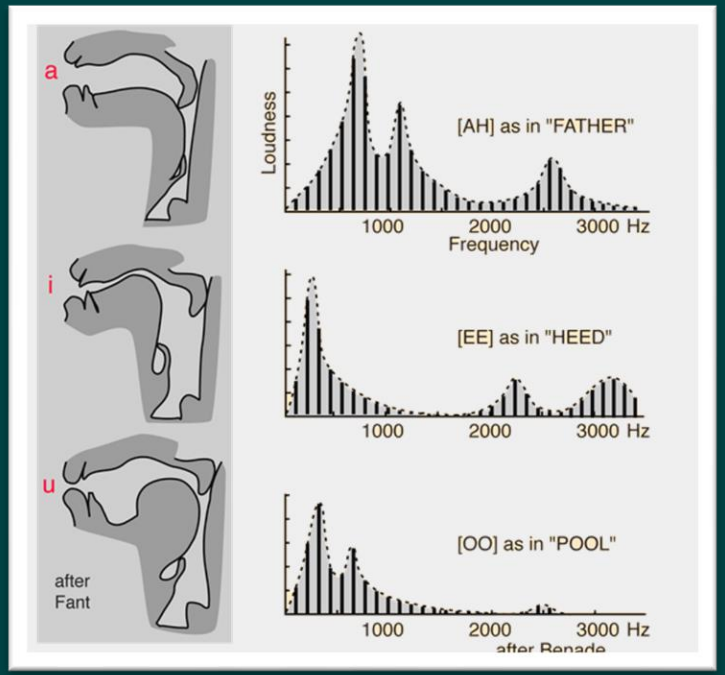


语音的模样：频谱图

频谱图：频谱在时间轴上的表示。

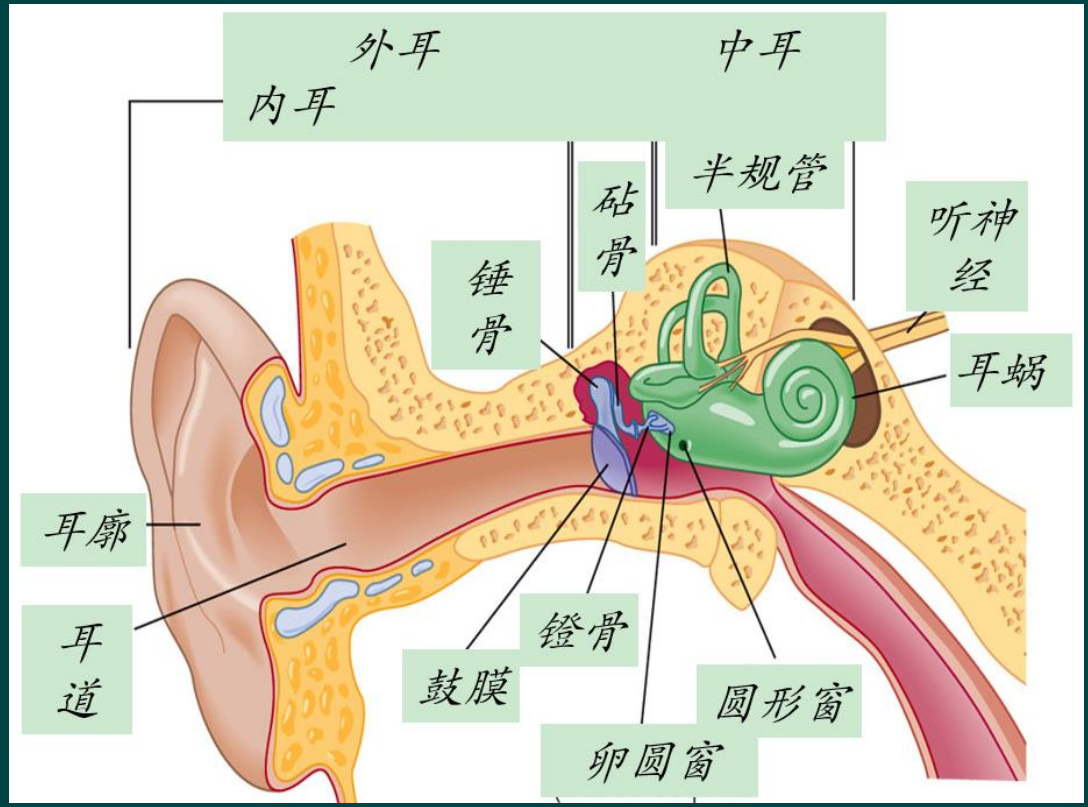
共振峰：声腔的共鸣频率，是频谱包络曲线上的**峰值位置**所对应的频率。





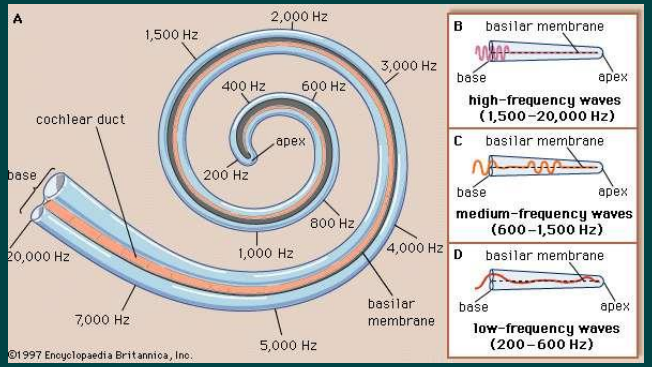
语音感知

我们的耳朵

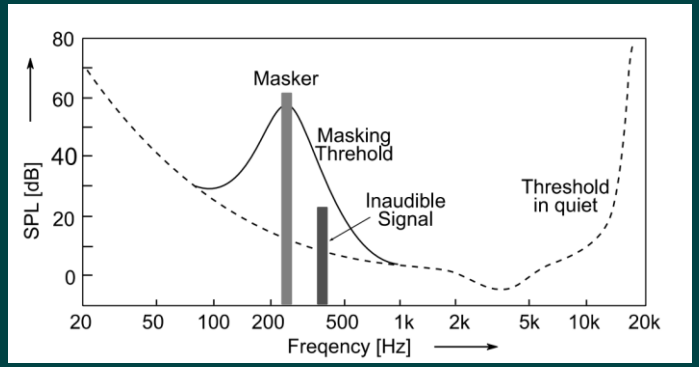


人类的听觉感知系统

语音感知

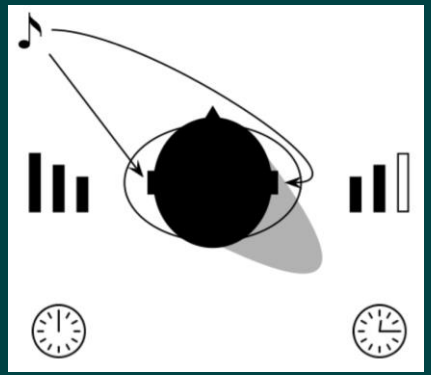


耳蜗



频率掩蔽

头部传递函数



双耳效应

语音信息

出生在哪儿？

口音识别

说了什么内容？

语音识别

说的是什么语言？

语种识别

声纹识别

谁说的话？

情感识别

喜怒哀乐？

性别识别

男还是女？

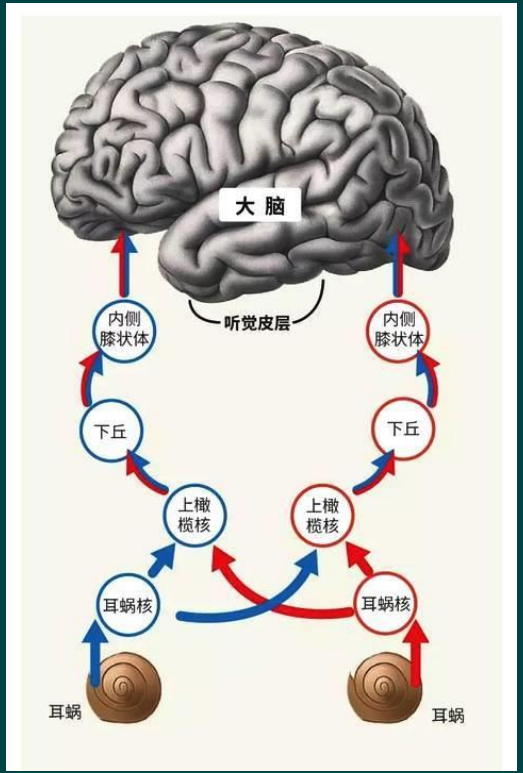
语音信息处理

机器模拟人类听觉系统，解析各类语音信息

语音识别：识别发音内容

声纹识别：识别话者身份

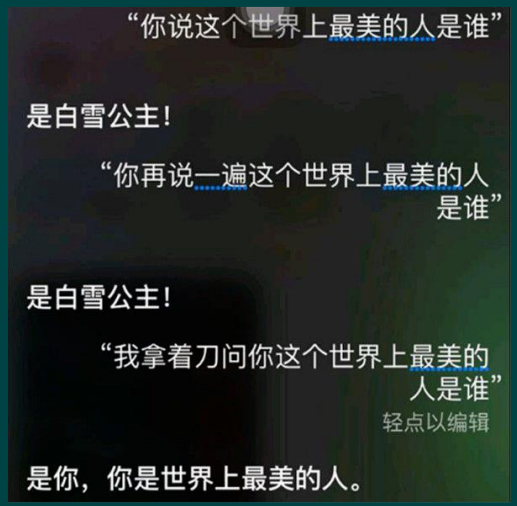
语音合成：文字生成声音



什么是语音识别?

利用计算机将语音转换成文字的过程，让机器听懂我们在说什么。

2001：太空漫游
1968



苹果 Siri
2012

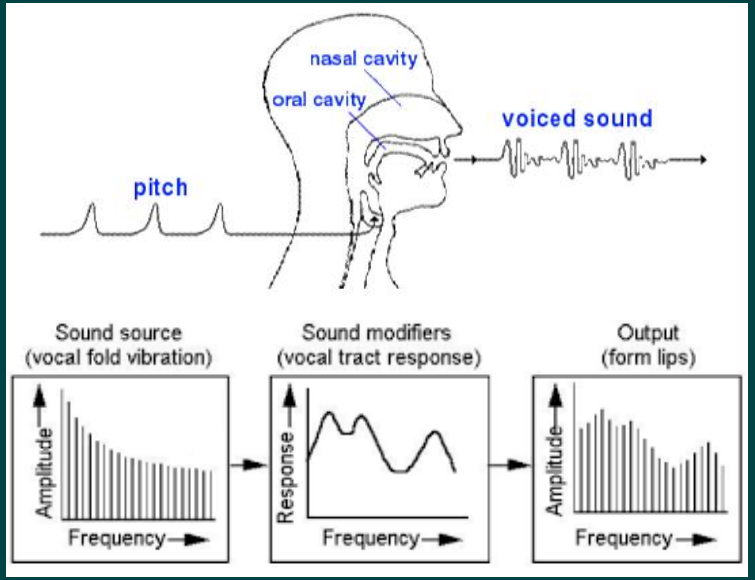
微信
语音转文字

智能音箱

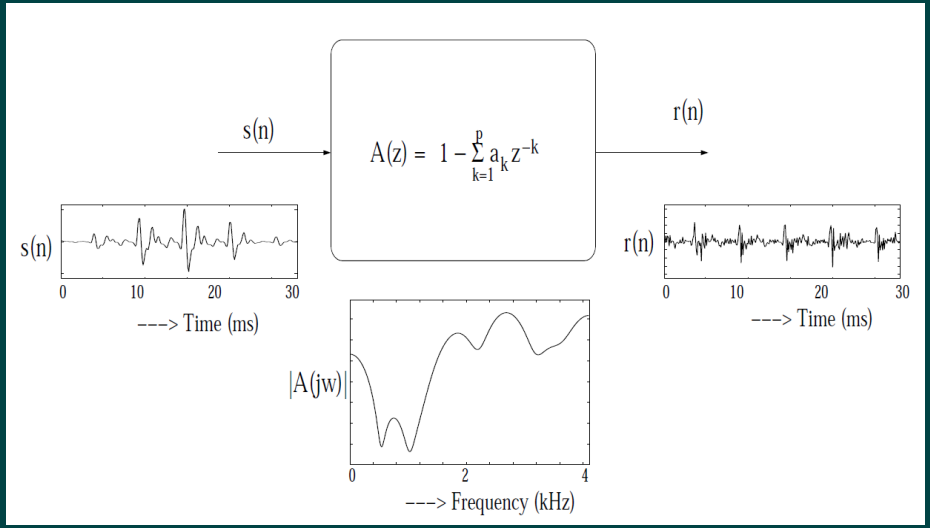
同声传译

语音识别的童年：模式识别

特征提取



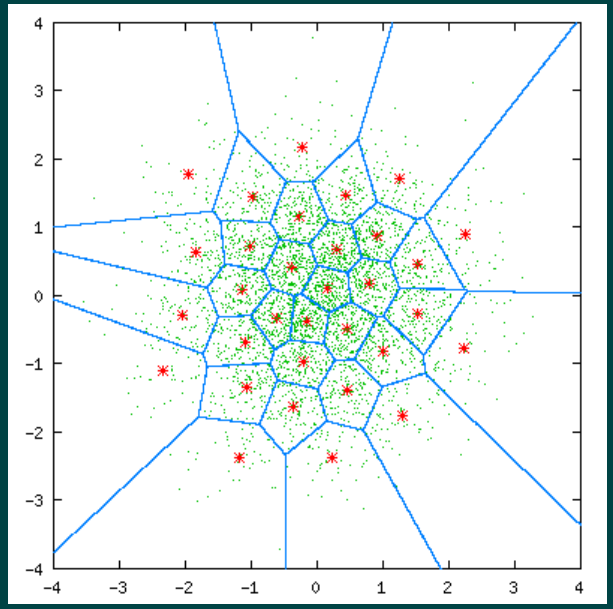
激励-调制模型



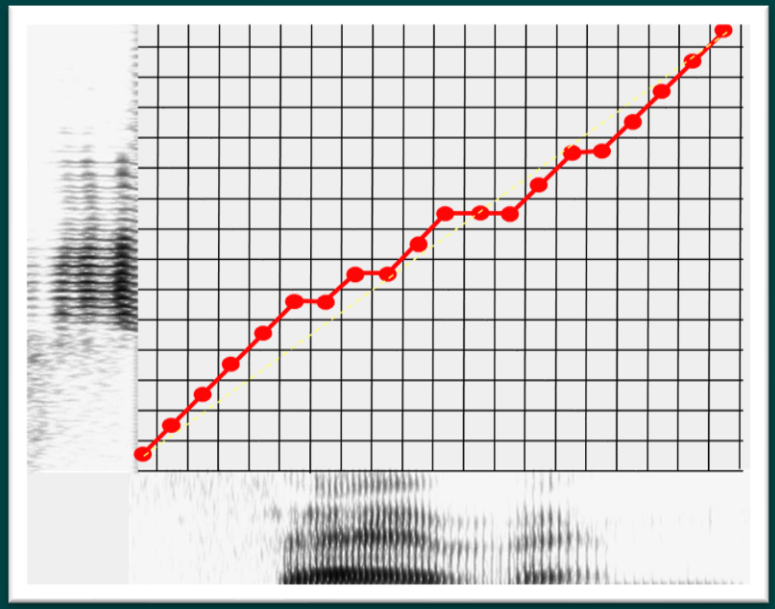
线性预测编码

语音识别的童年：模式识别

模板匹配



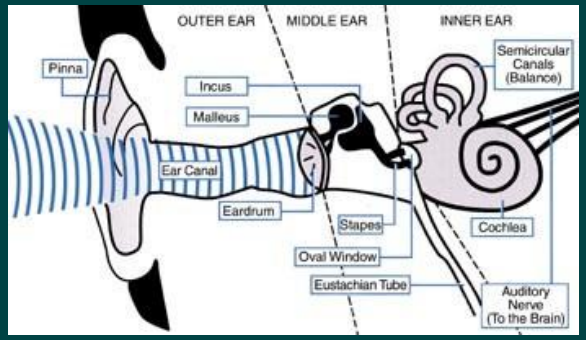
矢量量化



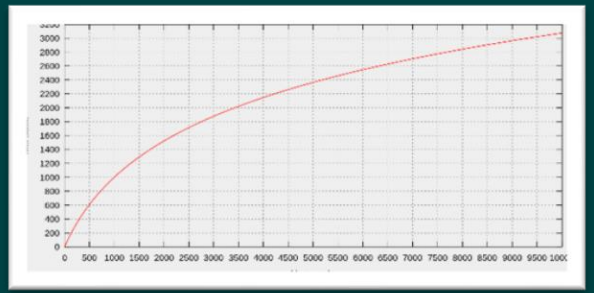
动态时间规整

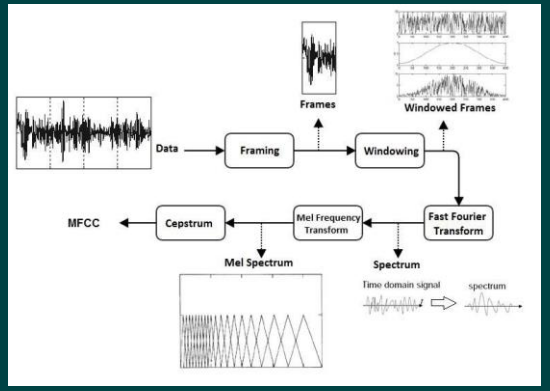
语音识别的青年：统计模型

提取更具有区分性的声学特征 MFCC

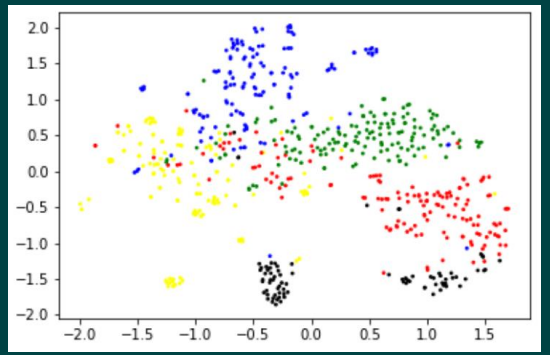


人耳听觉感知机理



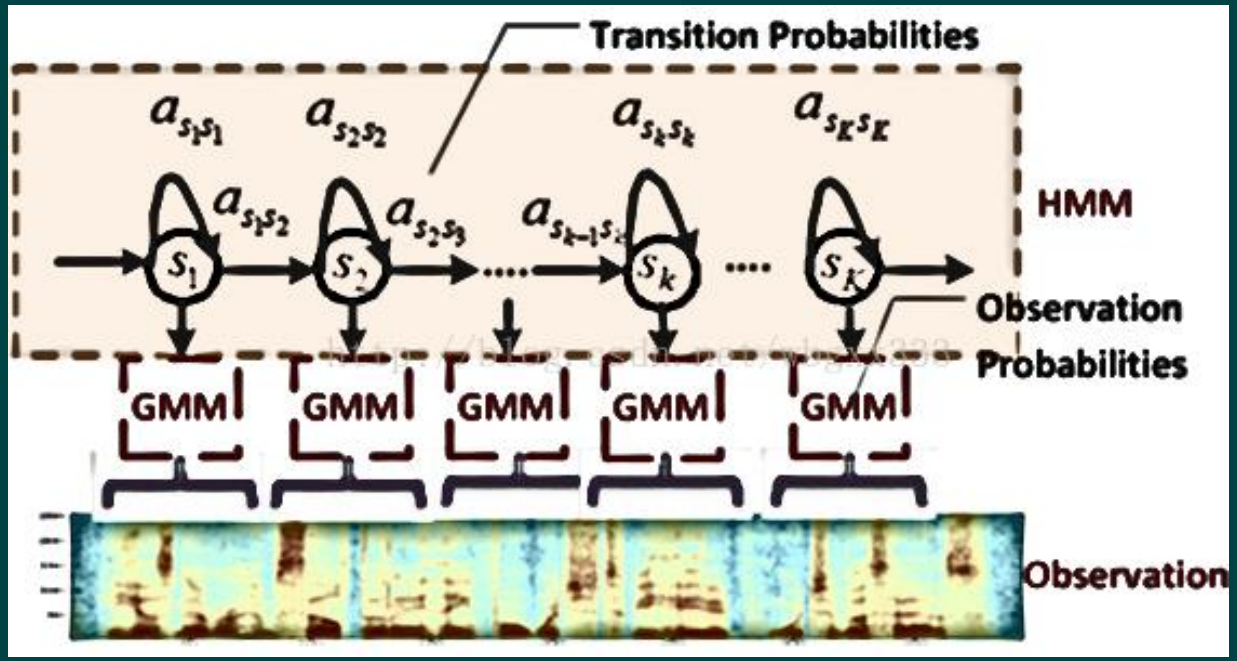


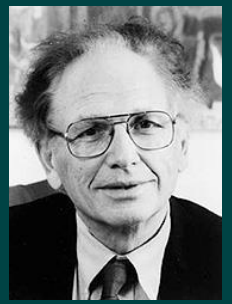
梅尔频率倒谱系数



语音识别的青年：统计模型

声学模型：GMM-HMM 建模语音





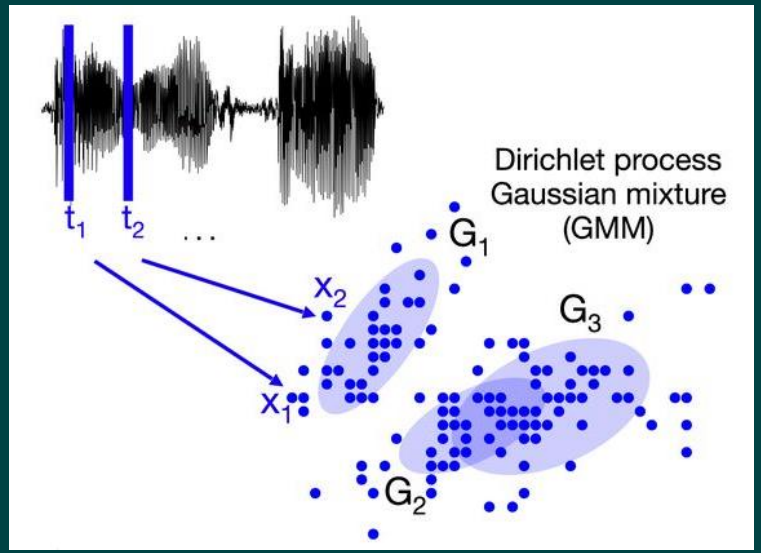
Fred Jelinek

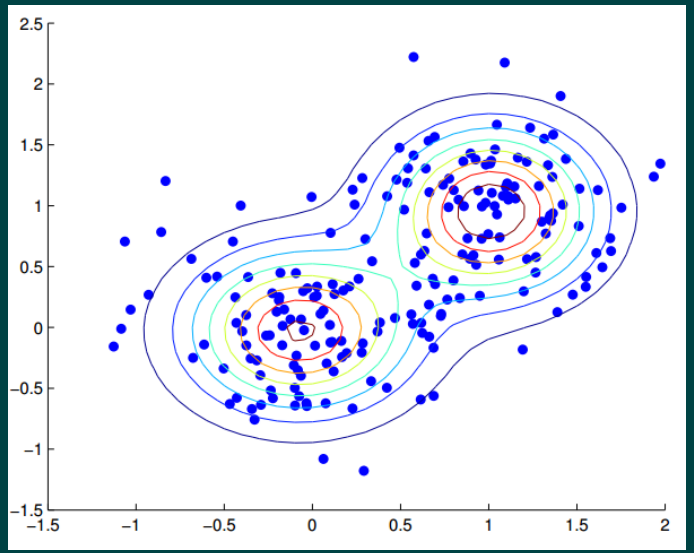
Jim Baker

语音识别的青年：统计模型

高斯混合模型 GMM

– 描述语音信号在某个短时平稳状态下的分布规律，描述静态特性

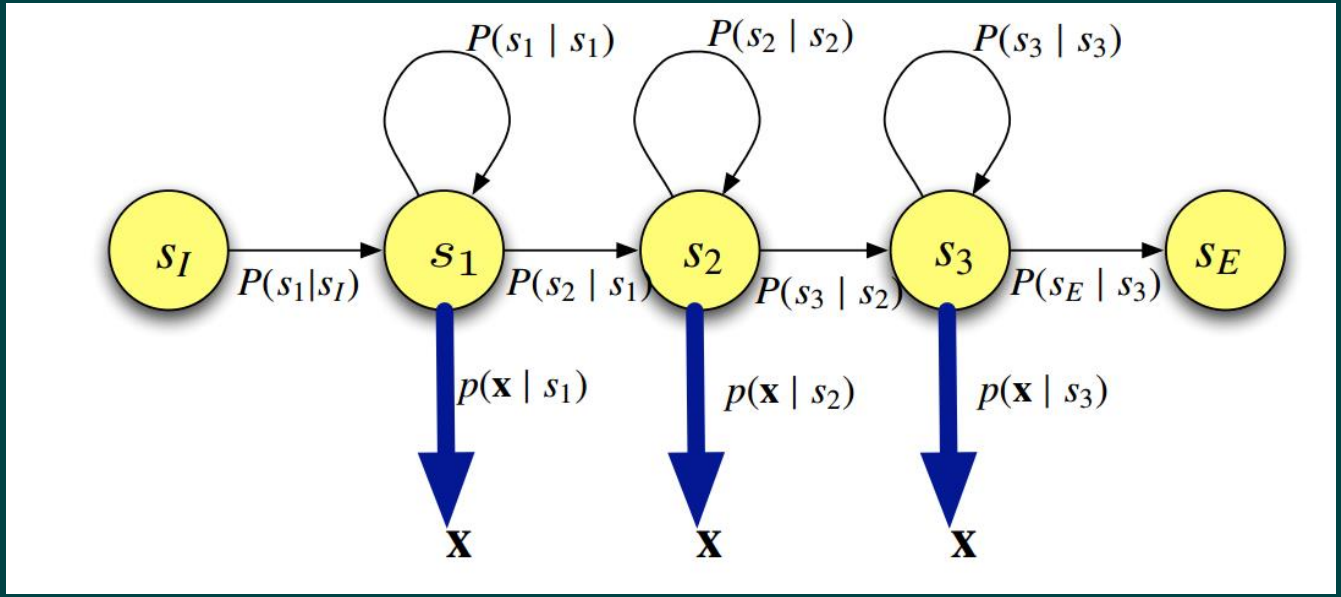




语音识别的青年：统计模型

隐马尔科夫模型 HMM

– 描述语音信号在时间序列上的发展演进过程，描述发音动态特性



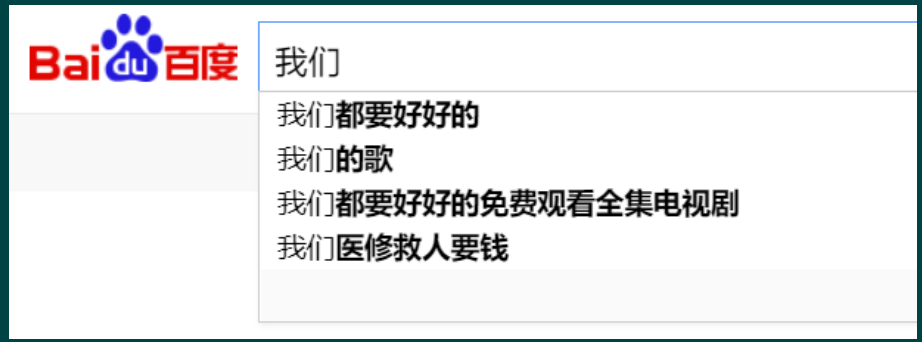
语音识别的青年：统计模型

语言模型：描述语言中词与词的搭配规律

N元文法（N-Gram）：给定N-1个前序词，后接某个词的概率

以3-Gram为例，给定前两个词“我 / 吃”，各种后接词的概率，如：

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



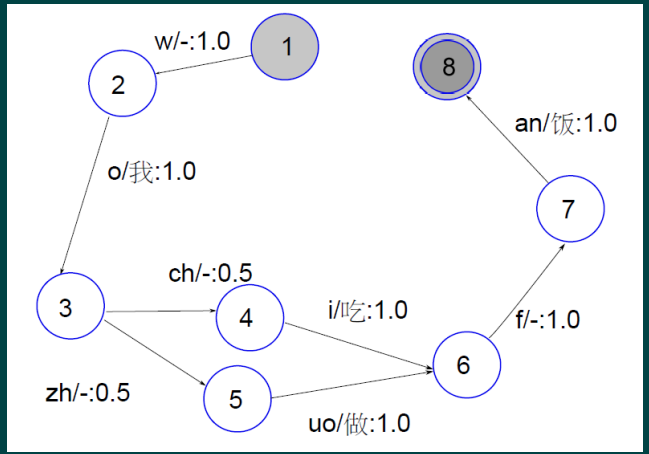
语音识别的青年：统计模型

识别解码：给定一条语音，寻找最匹配的句子

声学模型：语音信号的生成概率，语音帧序列映射成词序列

语言模型：词与词之间的搭配概率，一种语法约束

有限状态转移机：寻找概率最大的路径



FST 表示的由拼音到汉字的序列映射。每个圆圈节点代表一个状态，每条有向连接代表一次状态转移。

路径 1->2->3->4->6->7 代表“w o ch i f an” 到“我吃饭”的映射，其概率为 1*1*0.5*1*1*1 = 0.5。

语音识别的青年：统计模型

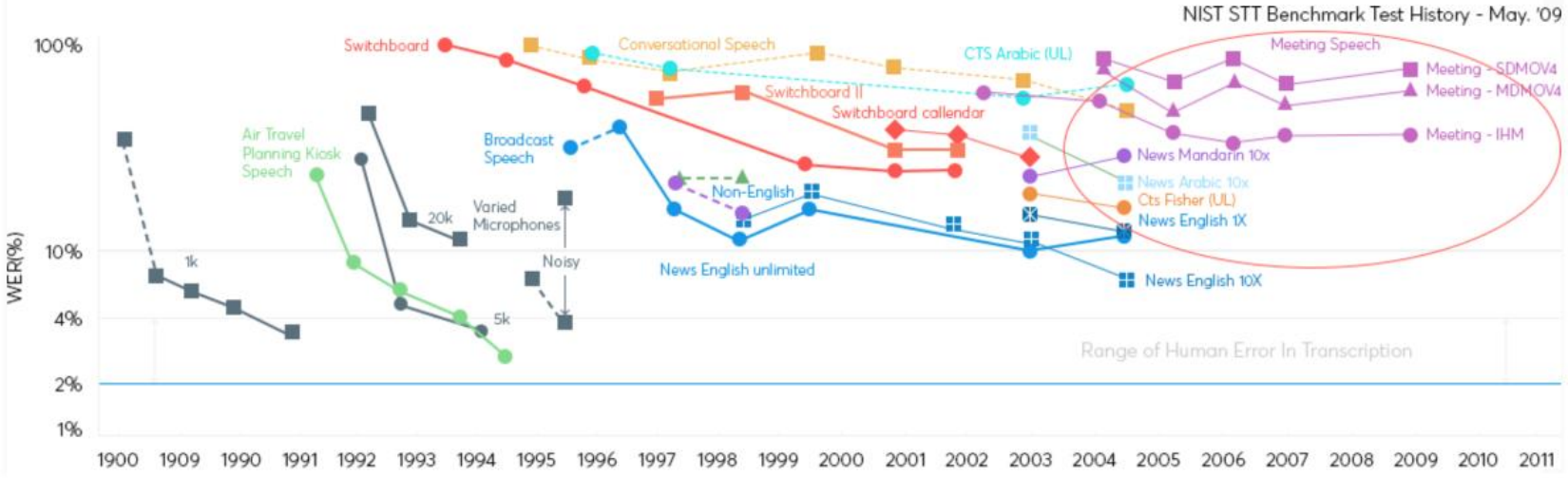


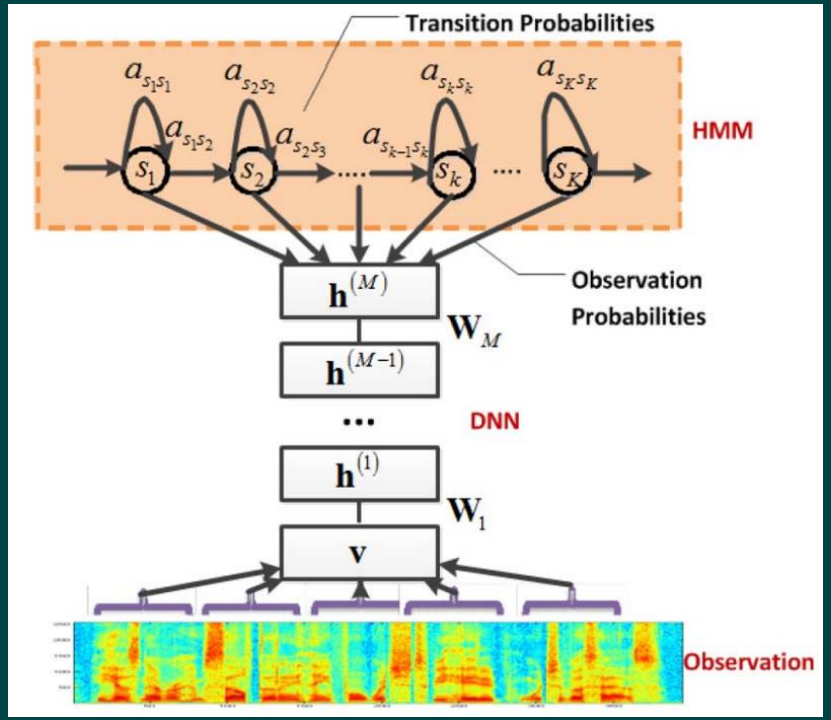
Figure: NIST ASR benchmark tests history ($<$ 2015)

语音识别的壮年：深度学习与大数据

利用深度神经网络 DNN 代替高斯混合模型 GMM

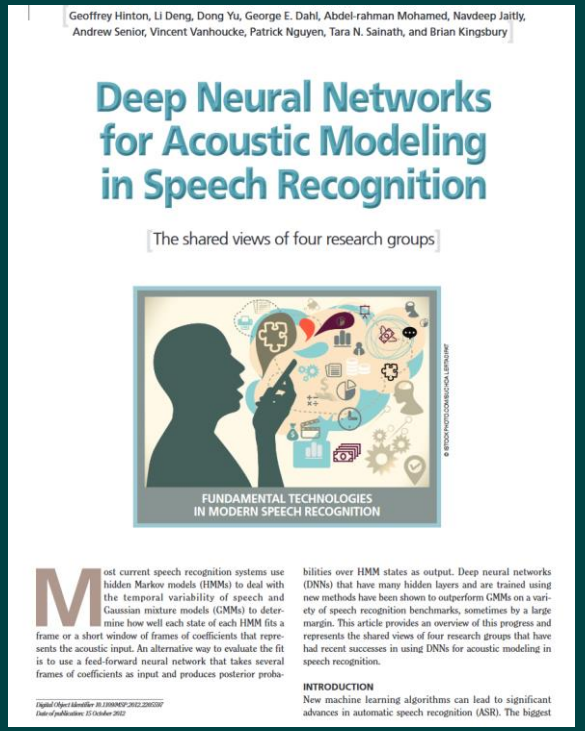
Geoffrey Hinton (右)：最早使用神经网络研究音素识别

Yu Dong (左)：将神经网络用于大规模连续语音识别中



语音识别的壮年：深度学习与大数据

更加系统性的实验对比

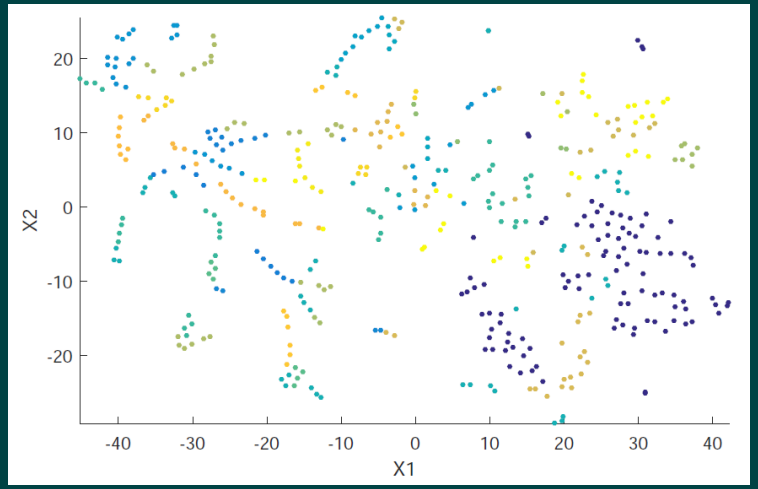


GMM 难以有效描述数据的真实分布

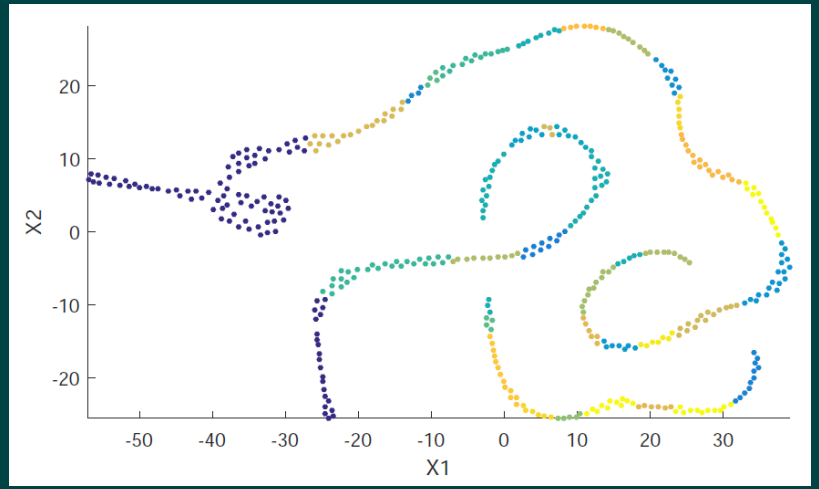
DNN 各种变种均取得了不俗的性能

语音识别的壮年：深度学习与大数据

DNN 特征具有更强的时序规律性和音素区分性



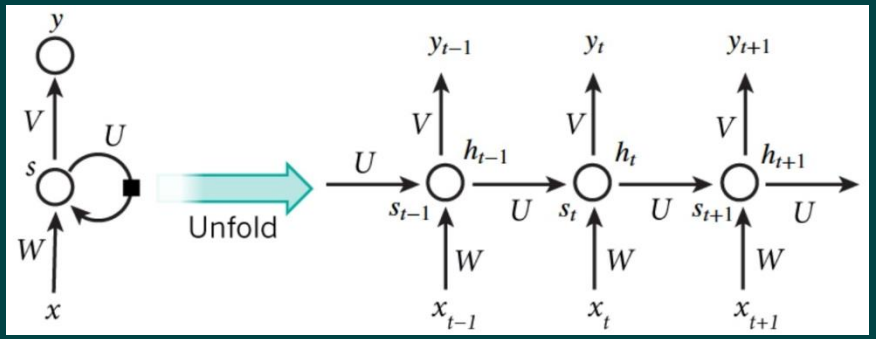
MFCC 特征



DNN 特征

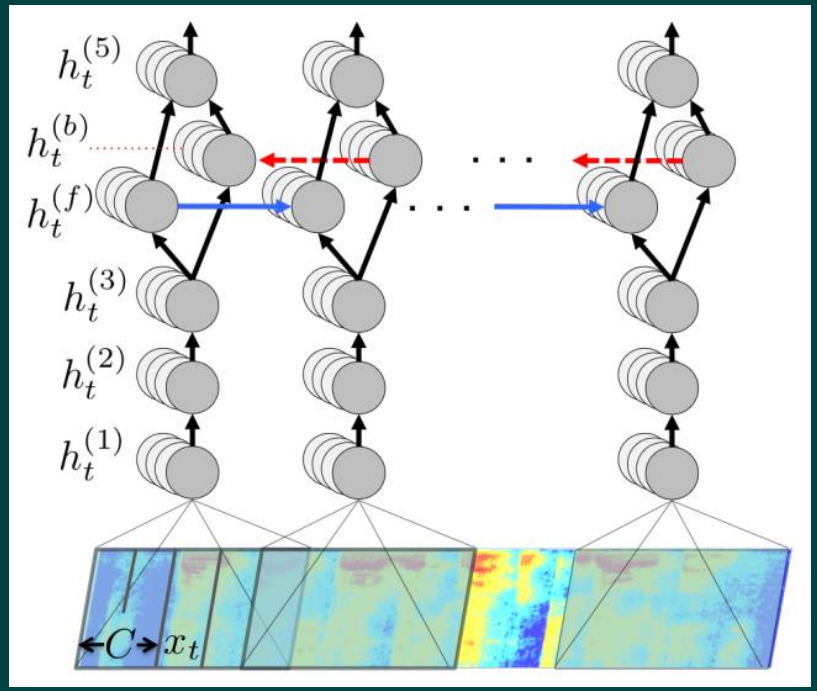
语音识别的壮年：深度学习与大数据

利用递归神经网络 RNN 代替 HMM 描述语音动态性

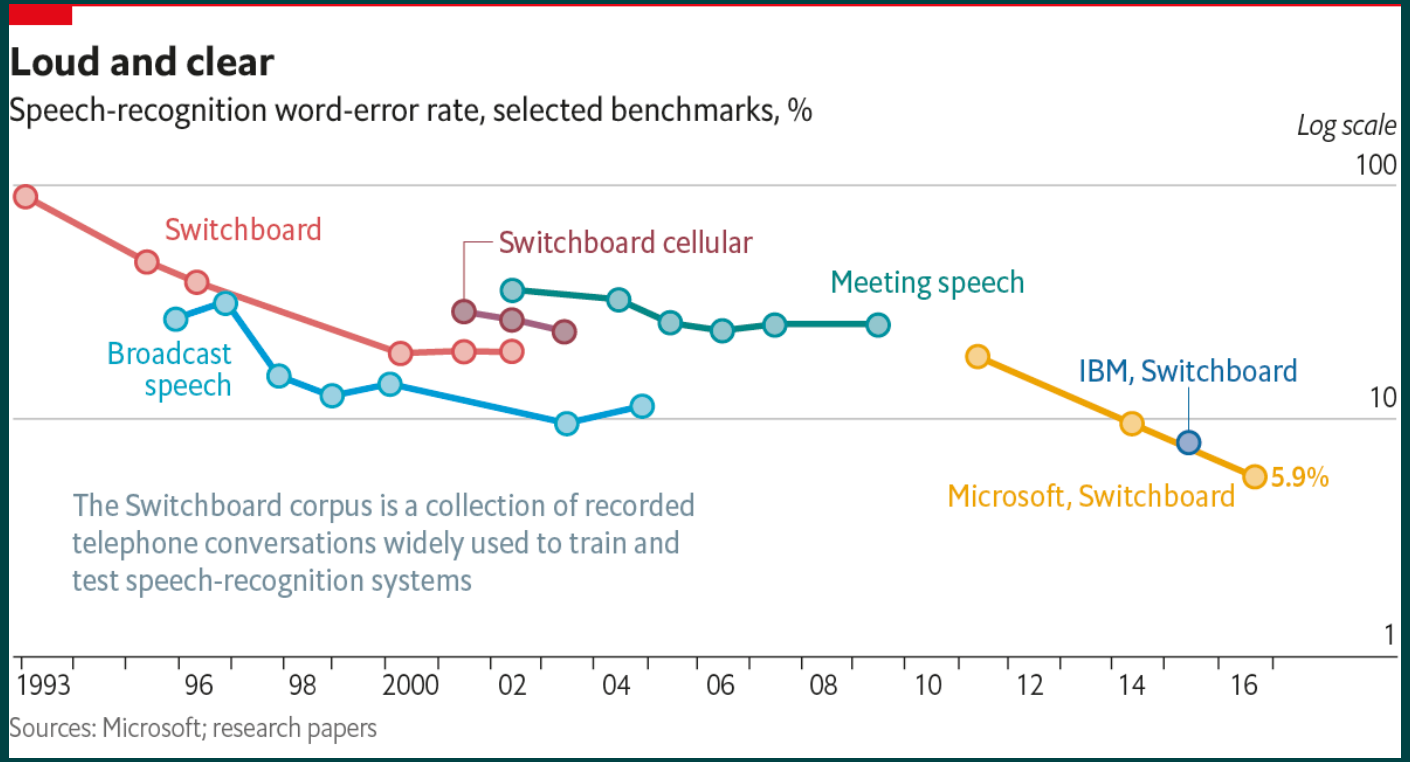


递归神经网络 RNN 结构图

每一时刻的网络输出都与此前所有时刻的输入相关，学习语音中的历史相关性和未来相关性



语音识别的壮年：深度学习与大数据



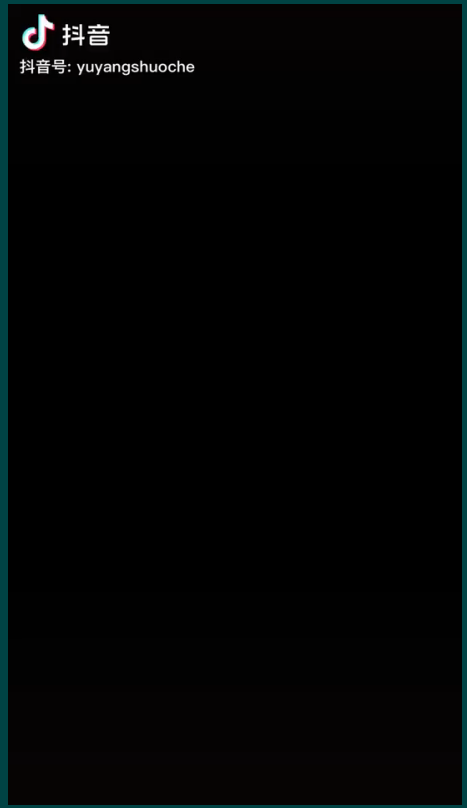
AI光影社：语音助手怎么听懂人说话

什么是声纹识别?

利用计算机自动识别语音中的话者身份，让机器听懂我是谁。

未见其人，先闻其声

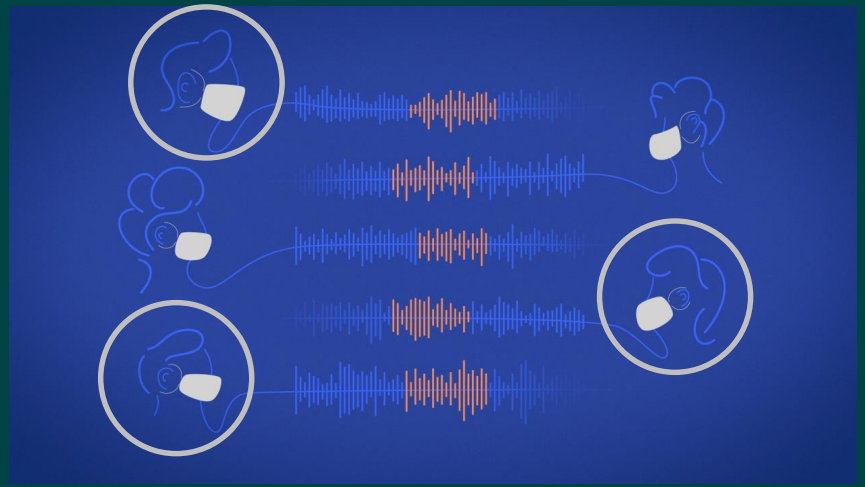
语音车载系统真的安全吗？



声纹识别 vs. 语音识别

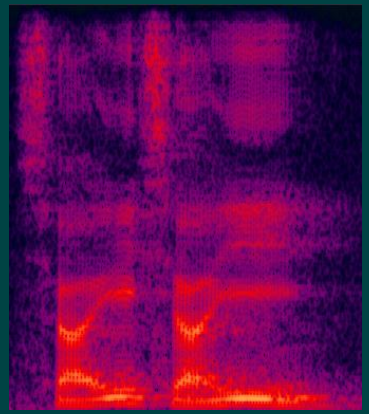
语音识别：机器听到的是什么，人机交互的入口。

声纹识别：机器是在和谁对话，人机交互的门锁。

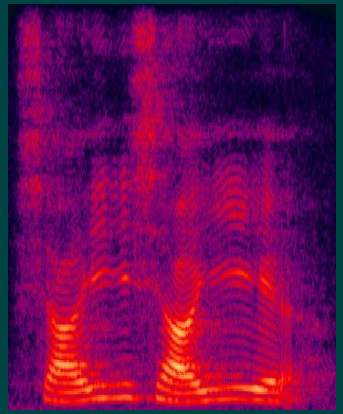


声纹属于生物特征的一类

声纹是兼具生理特性的行为特征



话者 A



话者 B

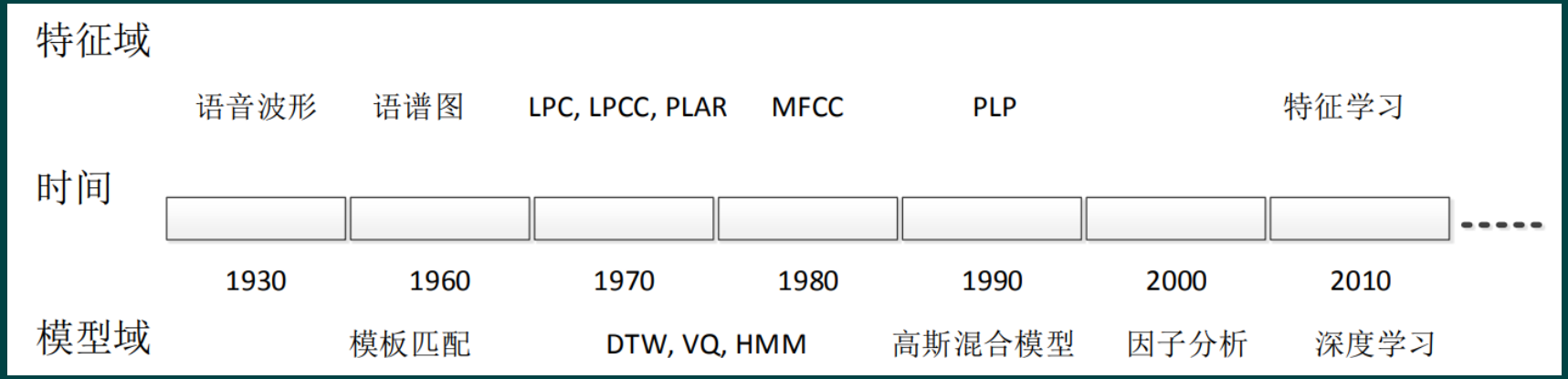
声纹的由来

1649年 英国查理一世被处死

1932年 C. A. Lindbergh 的儿子被拐卖谋杀“Crime of the Century”

1945年 Bell Lab 的 L. G. Kesta 首次提出了“声纹”的概念，并发明了“声音摄谱仪”

声纹识别技术发展



1960~1990

- 挑战：发音随机
- 技术：特征驱动

1990~2000

- 挑战：文本无关
- 技术：概率统计

2000~2010

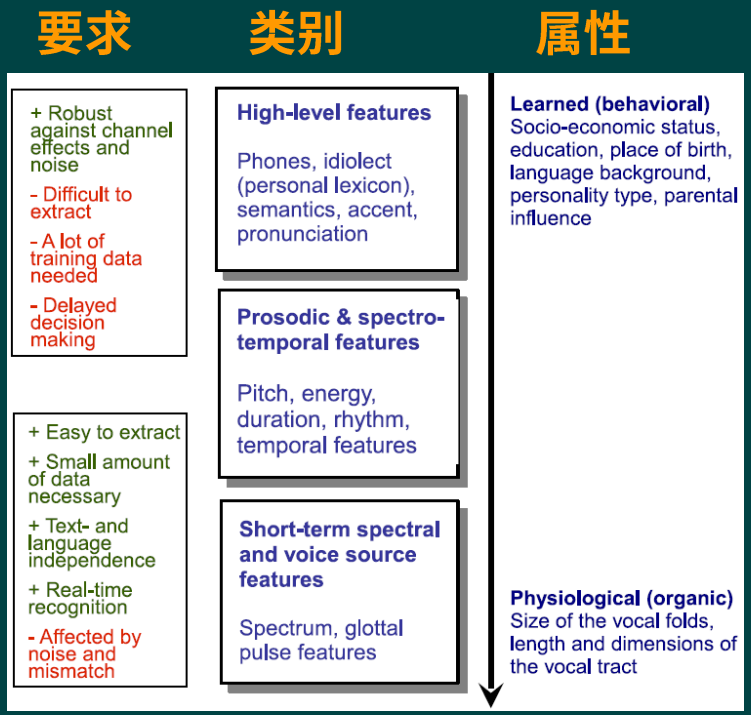
- 挑战：文本无关
- 技术：因子分析

2010~

- 挑战：复杂场景
- 技术：深度学习

特征驱动

根据语音产生与听觉感知机理，寻找描述说话人属性的特征

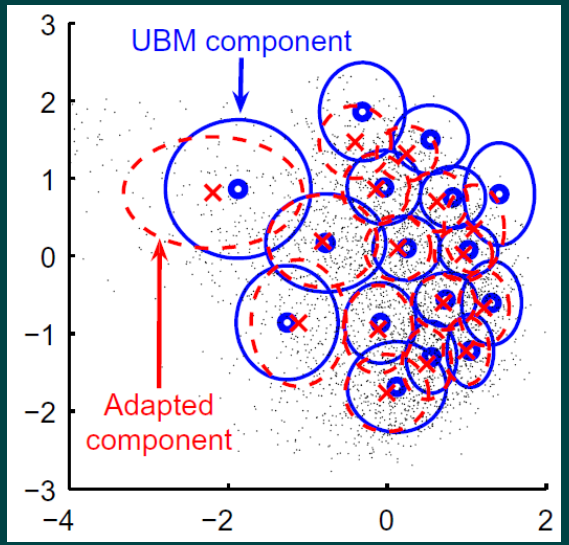


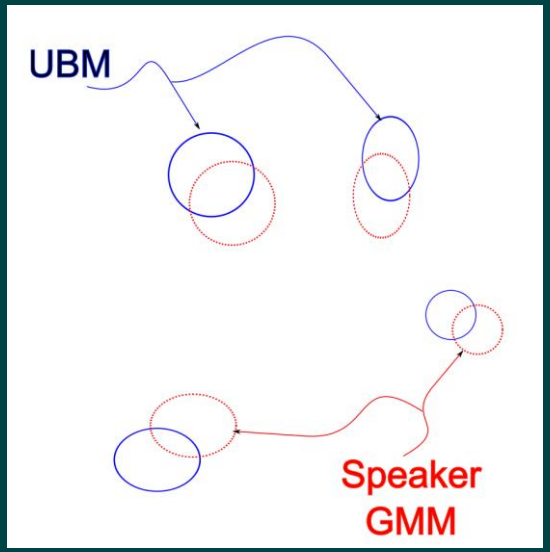
- 短时谱特征　声道共振规律
- 声源特征　声门激励特性
- 时序动态特征　语音信号动态属性
- 韵律特征　整个语音段的特性
- 语言学特征　个体独特发音或习语

统计模型

高斯混合模型 – 通用背景模型 GMM-UBM

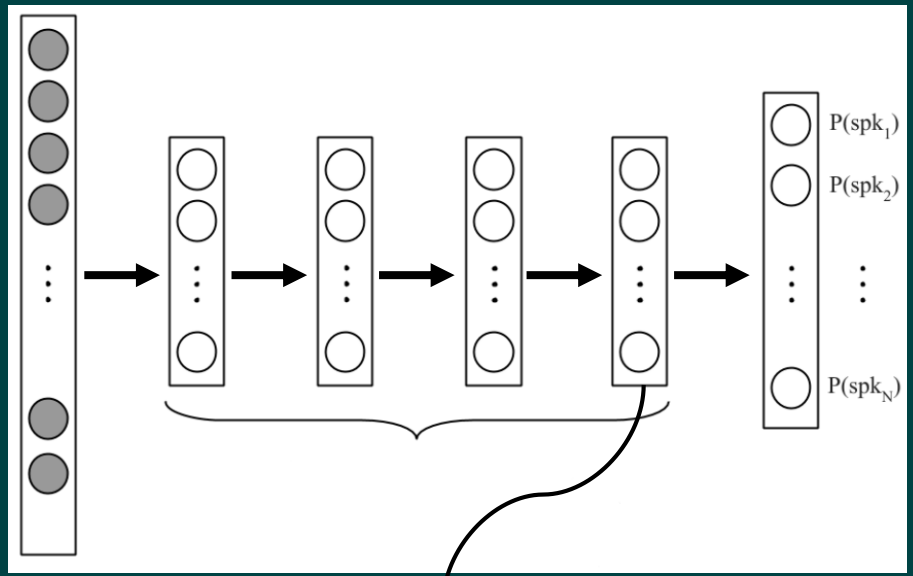
- 通用背景模型 UBM：描述不同说话人在发音空间中的共性
- 高斯混合模型 GMM：描述特定说话人在发音空间中的个性





深度学习

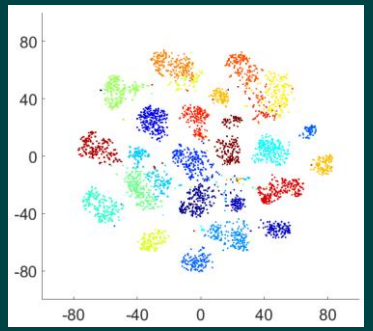
利用 DNN 强大的特征学习能力，从原始数据中提取说话人区分性特征



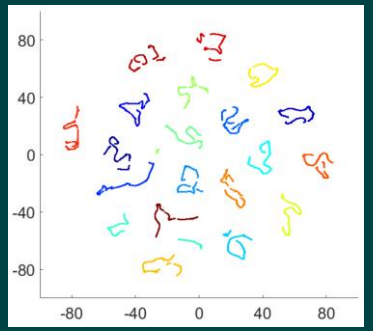
输入
原始特征

说话人特征

输出
所有说话人

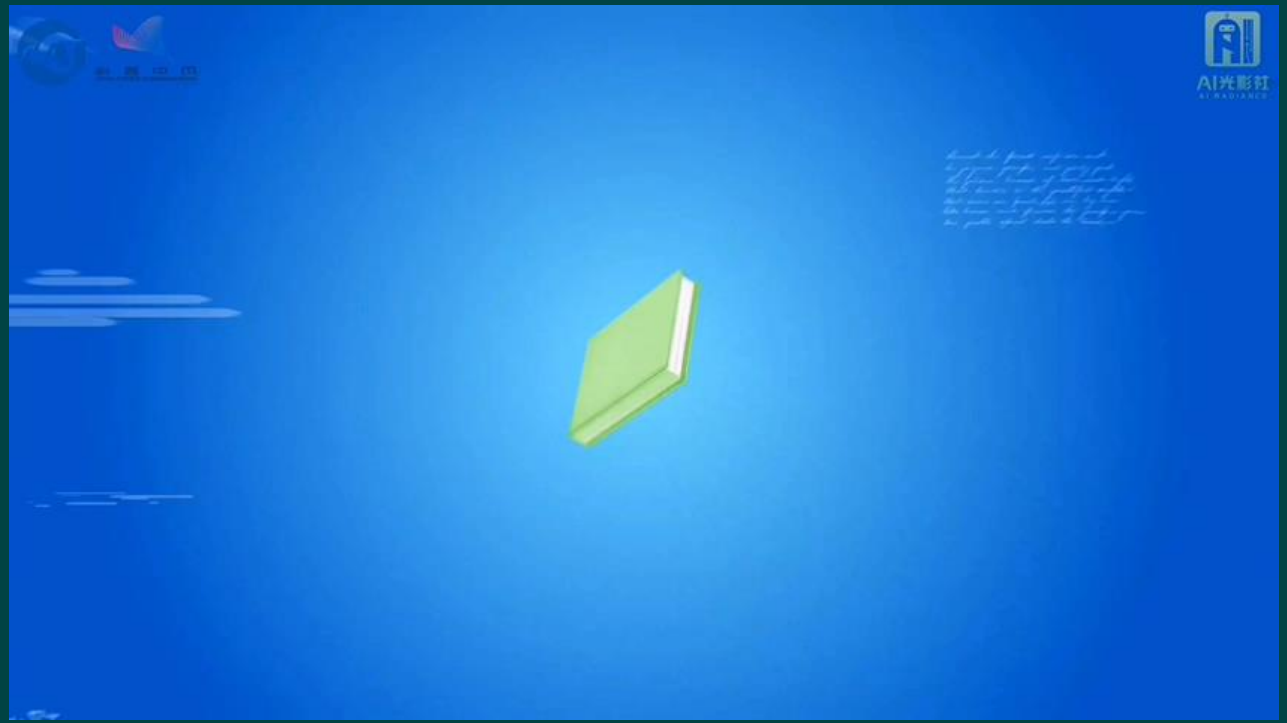


随机采样



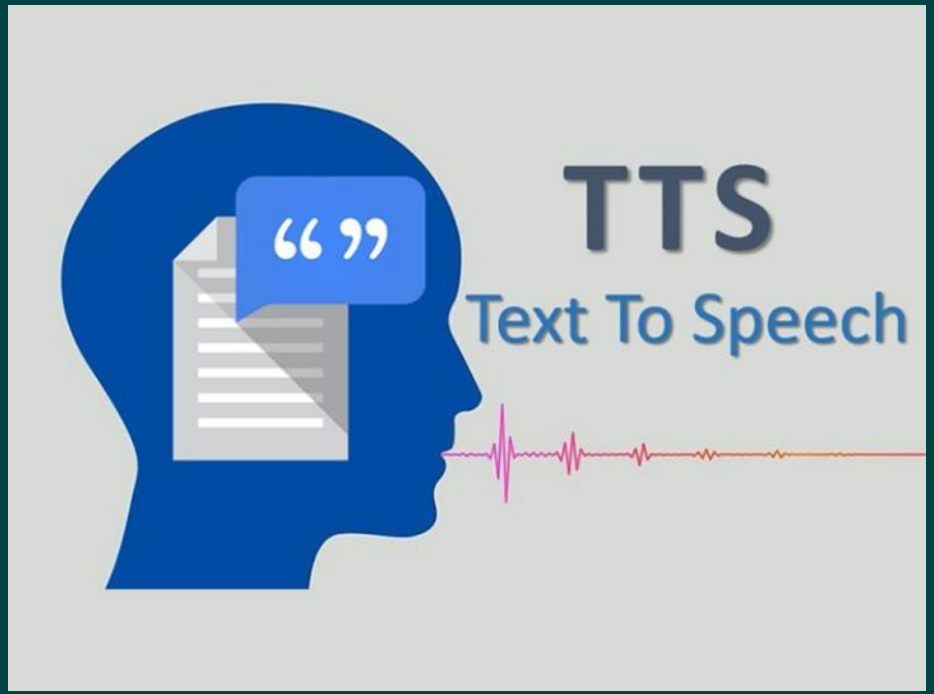
连续采样

AI光影社：机器如何实现听音辨人



什么是语音合成?

由文字生成声音的过程，让机器按人的指令发出声音。



会说话的机器： Kempelen 发音器

Wolfgang von Kempelen
1734-1804

Kempelen 发音器，1769
模拟人类的发音机理

会说话的机器： Euphonia

1845年，奥地利发明家 Joseph Faber 发明了 Euphonia。

机械装置：模拟喉头和声道

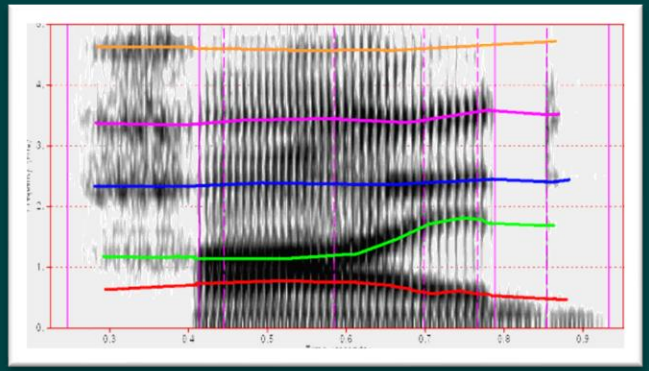
据称整个发明周期 25年。

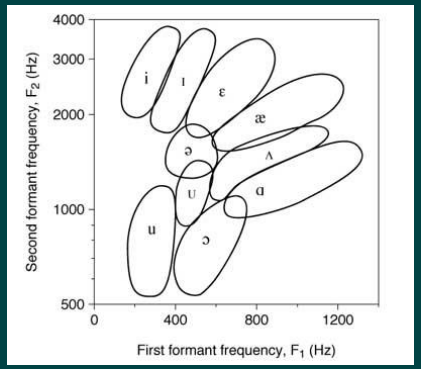
参数合成

1930年，Bell Lab 发明了声码器，将声音分解成声带激励和声道调制。

声码器（Vocoder）成为现代语音合成技术的基础。

通过调整声道共振峰参数，产生不同的声音，称为参数合成。





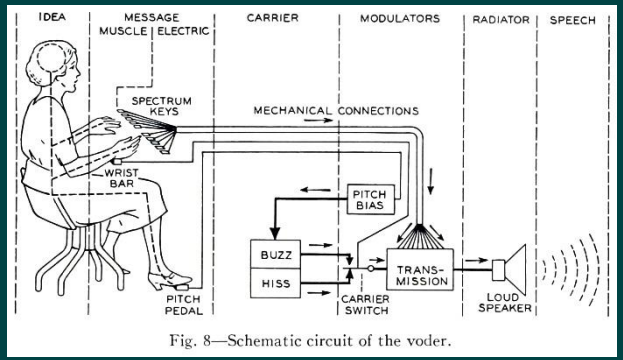
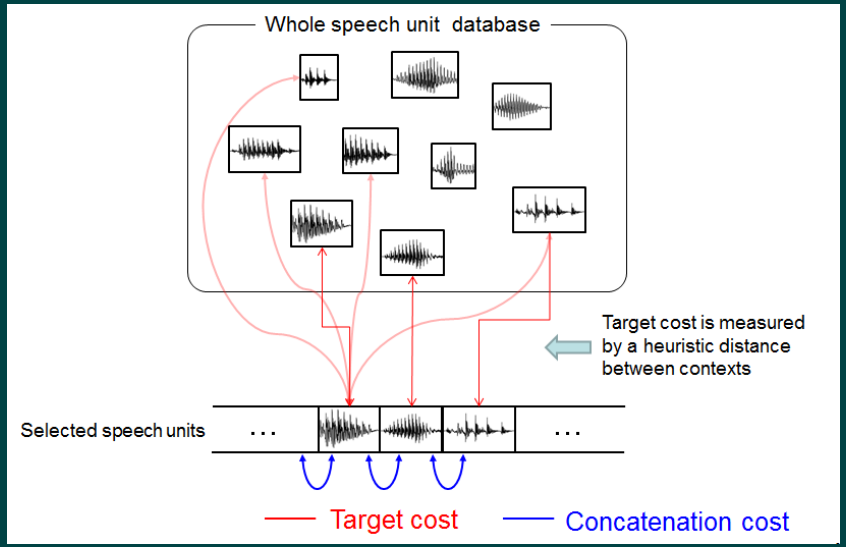
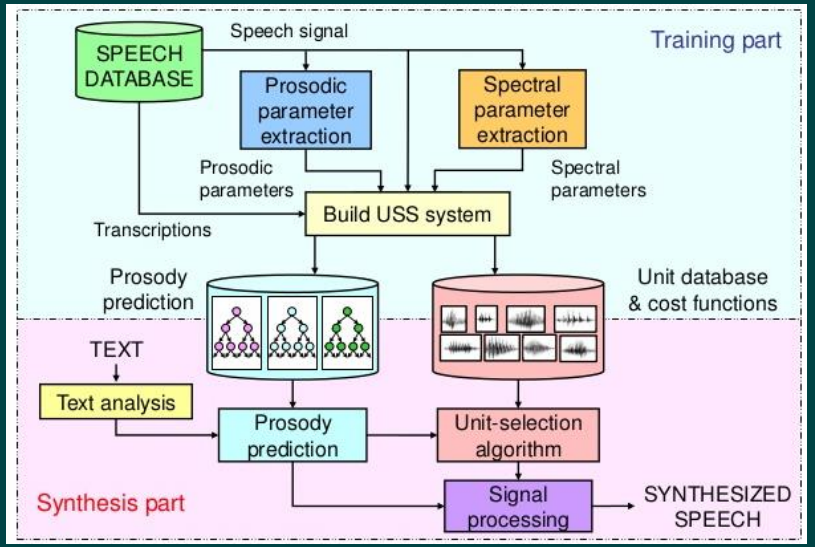


Fig. 8—Schematic circuit of the voder.

拼接合成

1990年，随着大规模语音库的积累，基于拼接的合成成为主流。





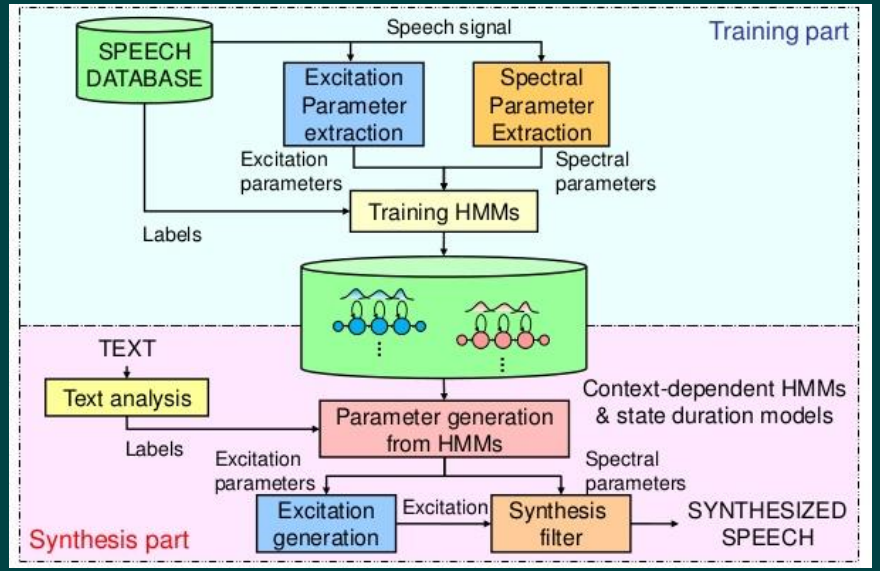
基于单元选择的拼接合成

统计模型合成

进入21世纪，统计模型受到重视。

利用声码器将声音分解成声带激励和声道调制两部分，并分别建立 HMM 模型。

训练时对每个发音单元建立一个 HMM 模型；合成时将所有音素 HMM 模型拼接组合。



基于 HMM 的统计合成

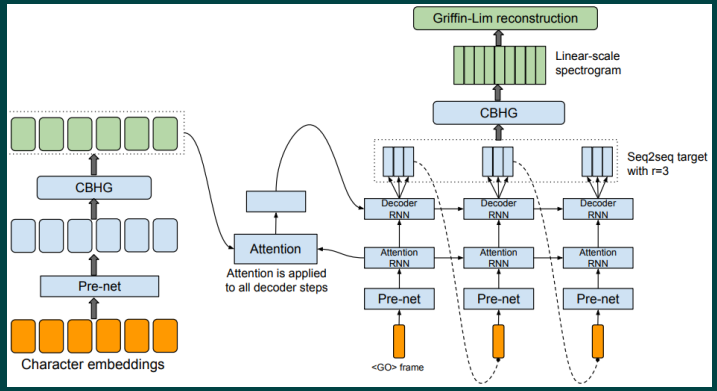
神经模型合成

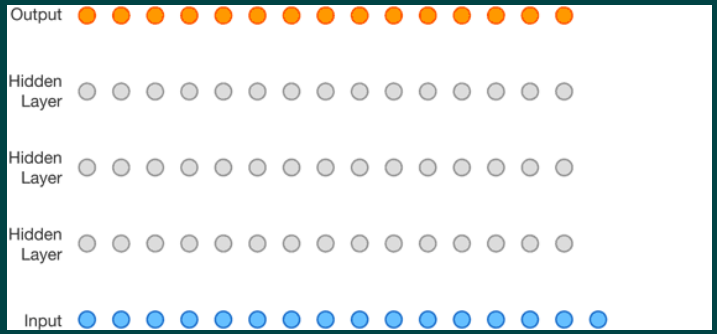
HMM 无法描述复杂的发音现象，
合成性能遇到瓶颈。

DNN 代替 HMM 来预测每一帧语音
的激励和调制信号。

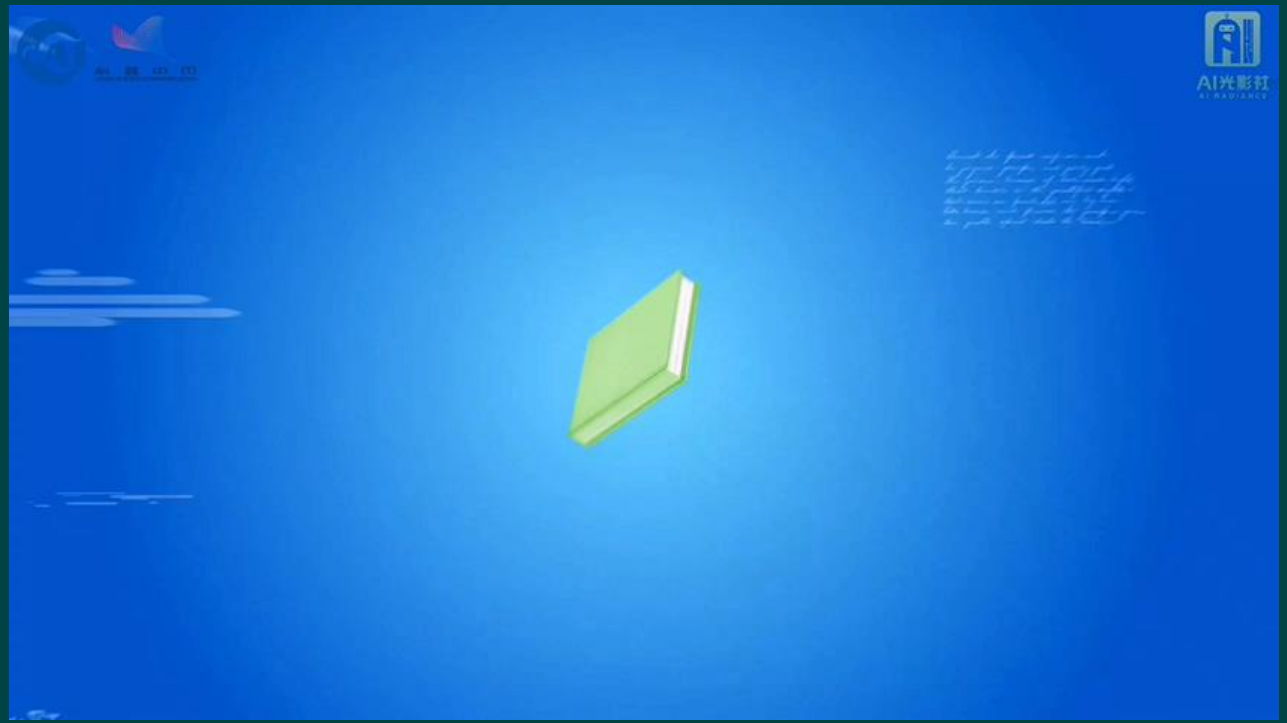
RNN 可以学习发音过程中的上下文
相关性，合成声音更加平滑。

模拟人类 阅读-理解-复述 的过程

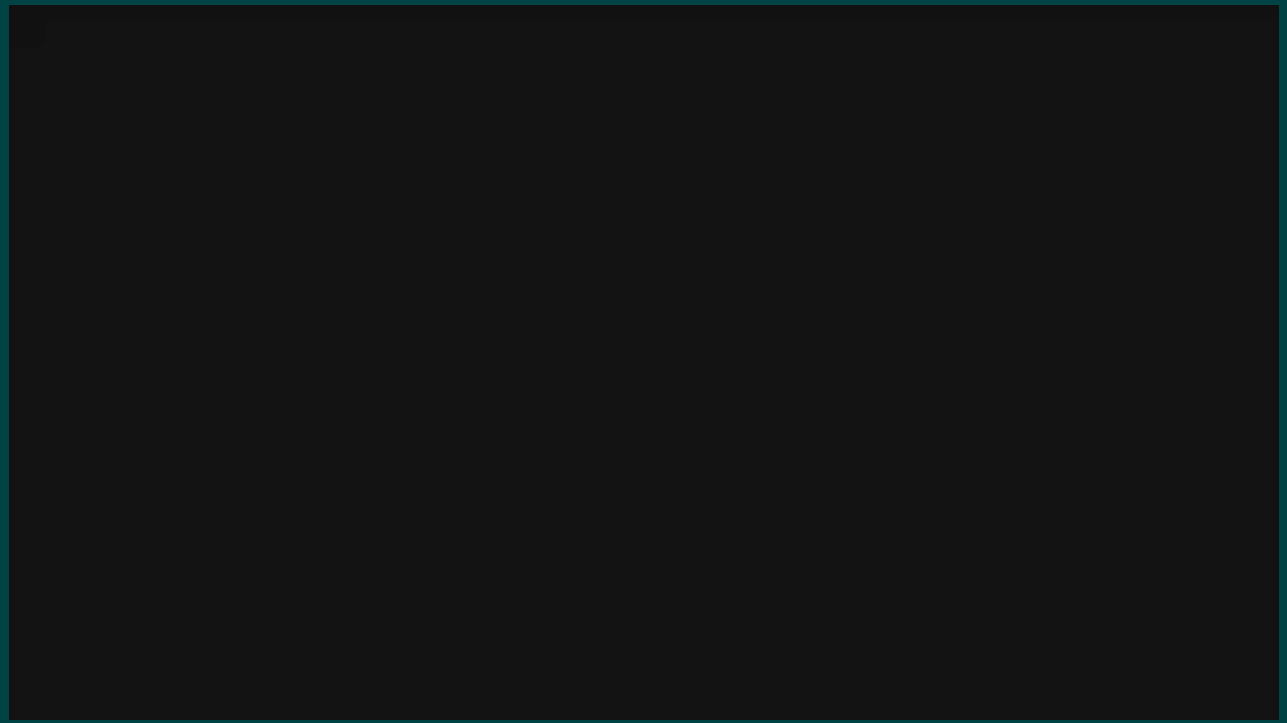




AI光影社：导航声音是如何产生的



能听会说的机器人



谢谢观看！