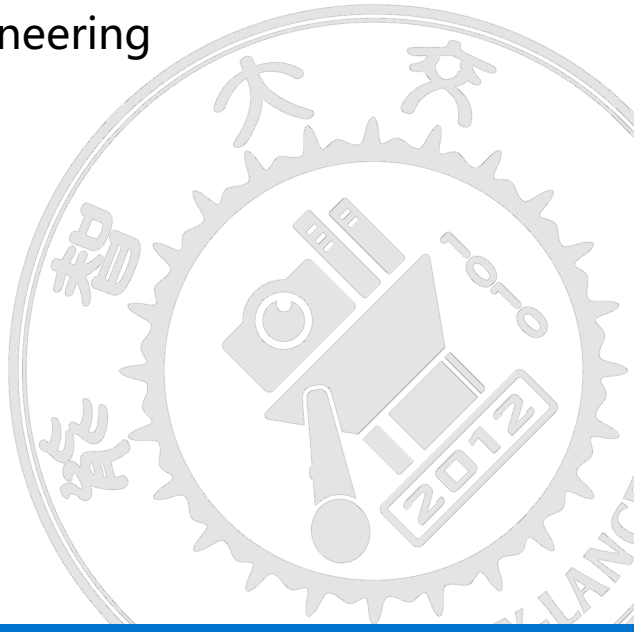
Lecture 0: 智能语音技术概述

Kai Yu and Yanmin Qian

Cross Media Language Intelligence Lab (X-LANCE)
Department of Computer Science & Engineering
Shanghai Jiao Tong University

2021

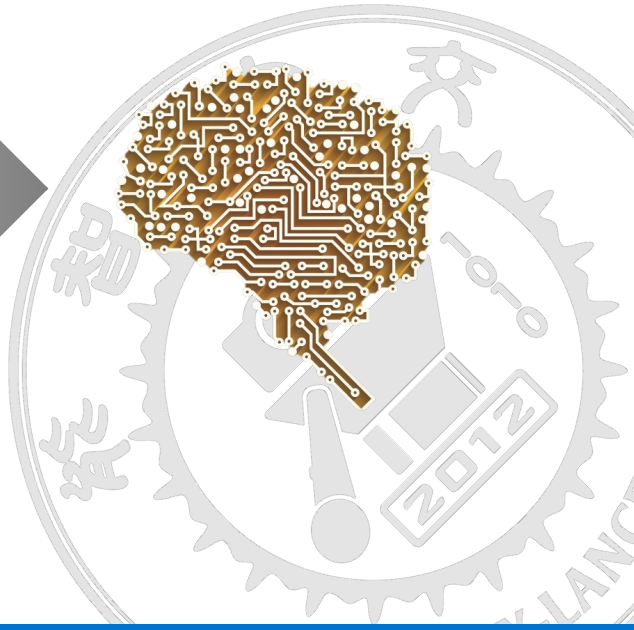


跨媒体语言智能 = **智能语音** + 自然语言 + 图像语义

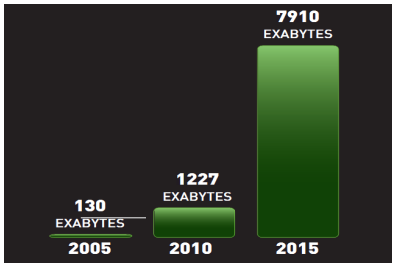
CROSS-media **LAN**guage intelligen**CE** = **X-LAN**CE

人工好理解，什么是智能？

计算	感知/表达	认知	知识处理
存储、计算	听、说、看 闻、触、行动	理解、思考 反馈、适应	分析、推理 演绎、归纳

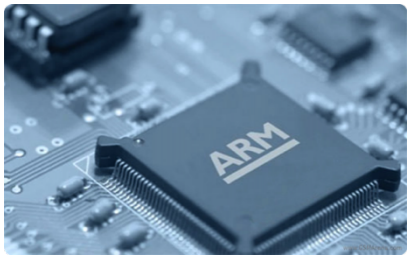


移动互联网促成了人工智能四大基础、三大爆点

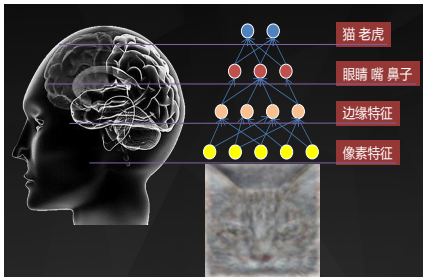


大数据

云计算



超级硬件



深度学习

大数据分析

实体机器人

人机交互

智能口语对话是物联网时代的人工智能入口

PC互联网（0.1B）

手机无线网（1B）

硬件物联网（10B）

文本搜索

口语/文本对话

口语对话

智能口语对话是物联网时代的人工智能入口

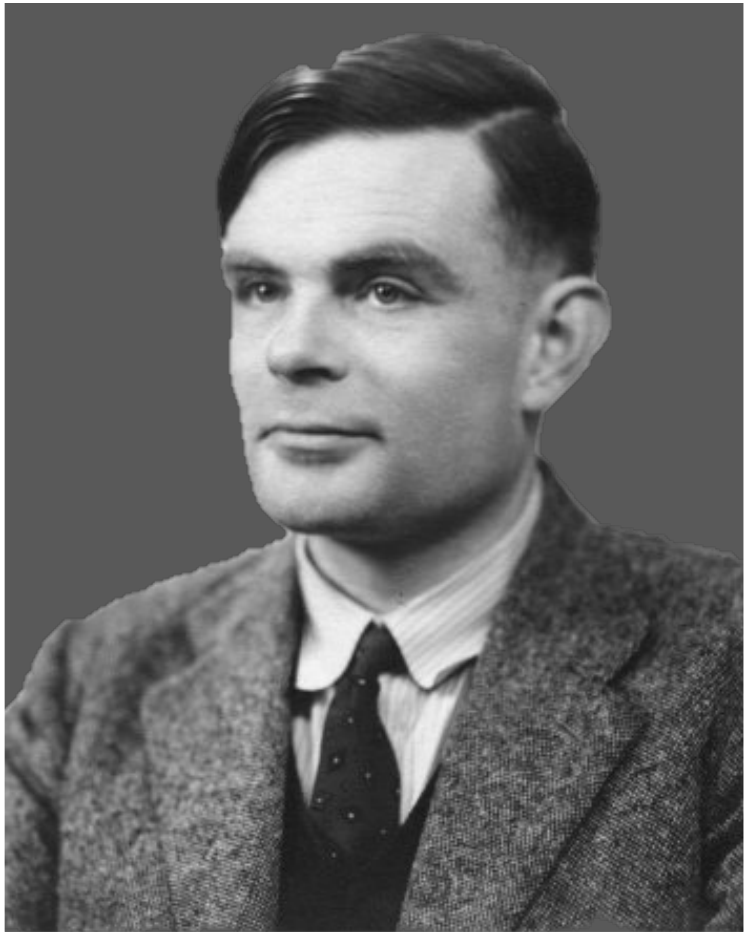
Amazon Echo
(2014)

Microsoft Button
(2015)

Google Home
(2016)

Apple Pod
(2017)

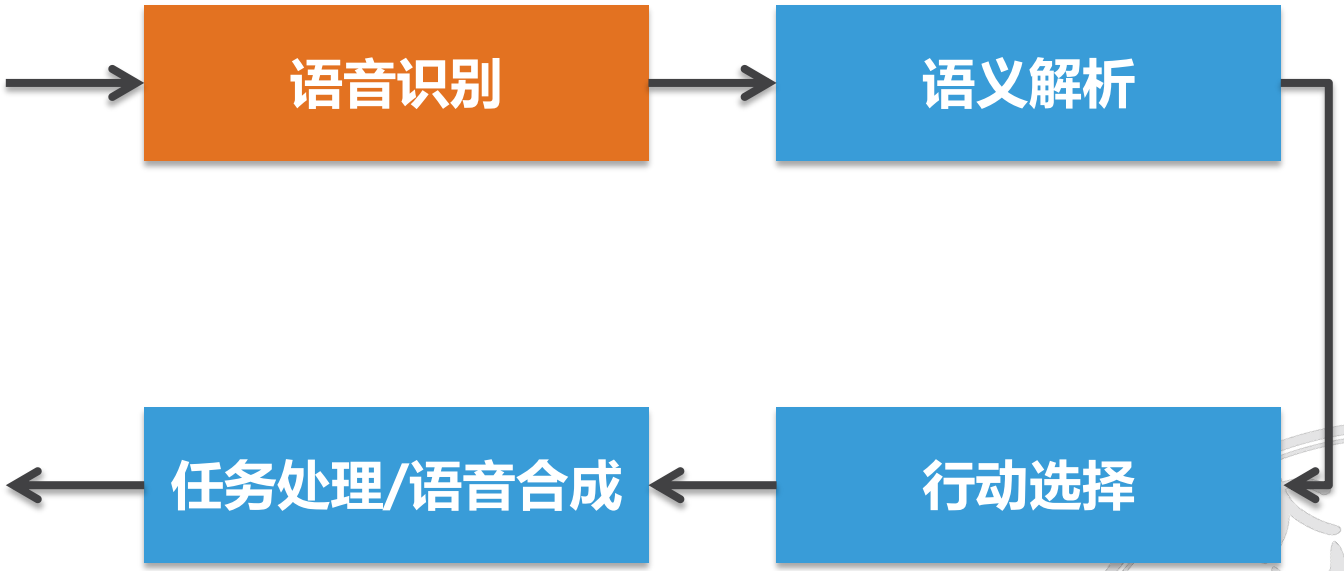
语言及对话式人工智能



语言智能 是
人工智能的皇冠

对话智能 是
人工智能皇冠上的明珠

经典人机口语对话系统架构



观点：**语音**就是**键盘替代品**，**单轮交互**，信息**一次性处理**

语音识别

输入——

- 音频:
  - 一个整数或实数序列
  - 16000 次采样/秒 (16K Hz)
  - $\boldsymbol{x} = \{x_1, \dots, x_T\}$
- 特征提取:
  - 25ms 窗宽
  - 10ms 帧移
  - 一个**实数向量**序列
  - $\boldsymbol{O} = \{\boldsymbol{o}_1, \dots, \boldsymbol{o}_T\}$

输出——

- 词序列
  - $\boldsymbol{w} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_m\}$
  - $T \gg m$

难点

- 变长序列（音频/词均变长）
- 词表数量巨大
- 环境、说话人干扰

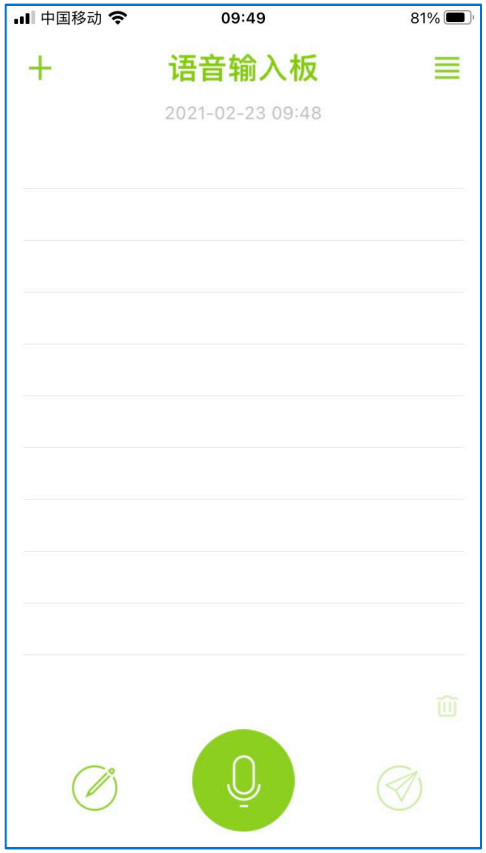
机器学习类型

- 孤立词：分类
- 关键词：序列标注
- 连续语音识别：
  序列分段+分类+搜索

语音识别错误率有多高？

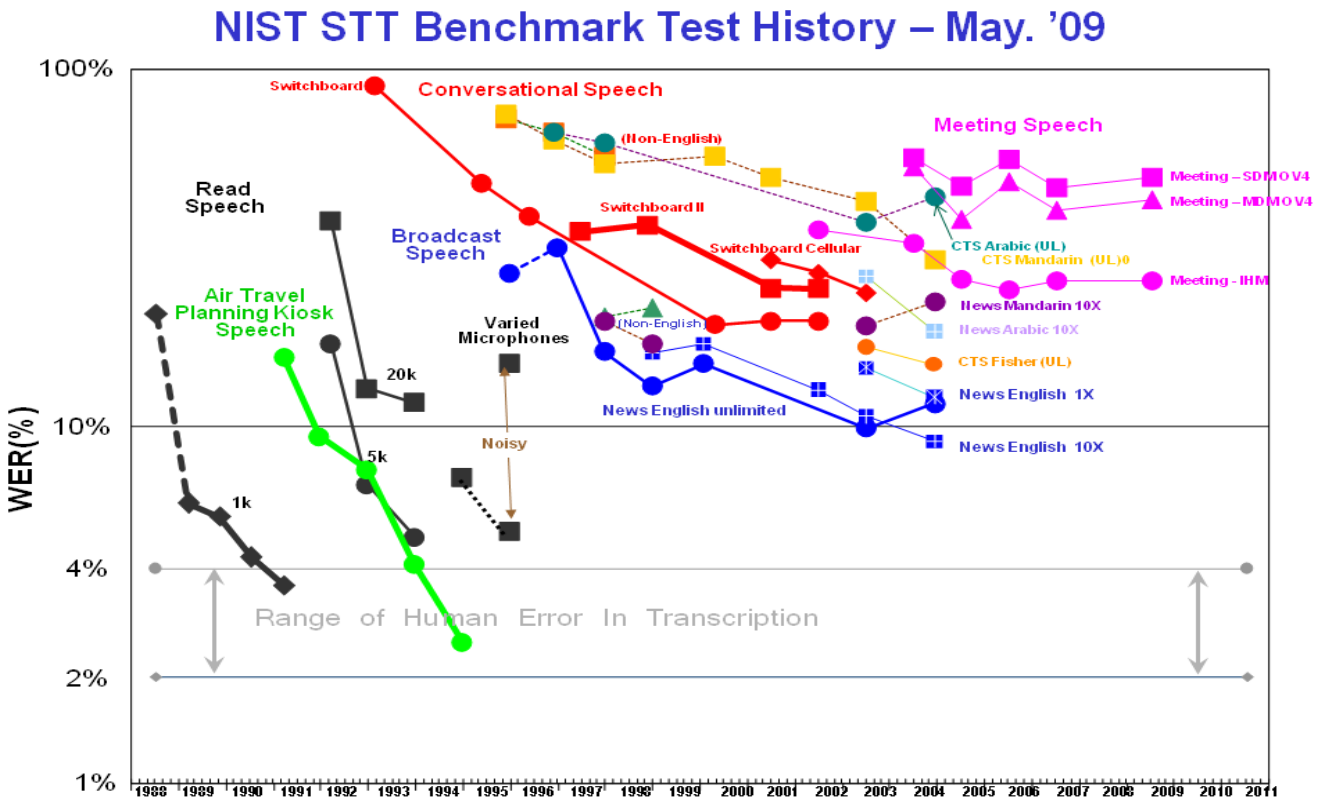
思必驰 AISPEECH

思必驰-上海交通大学
智能人机交互联合实验室



语音识别错误率有多高？

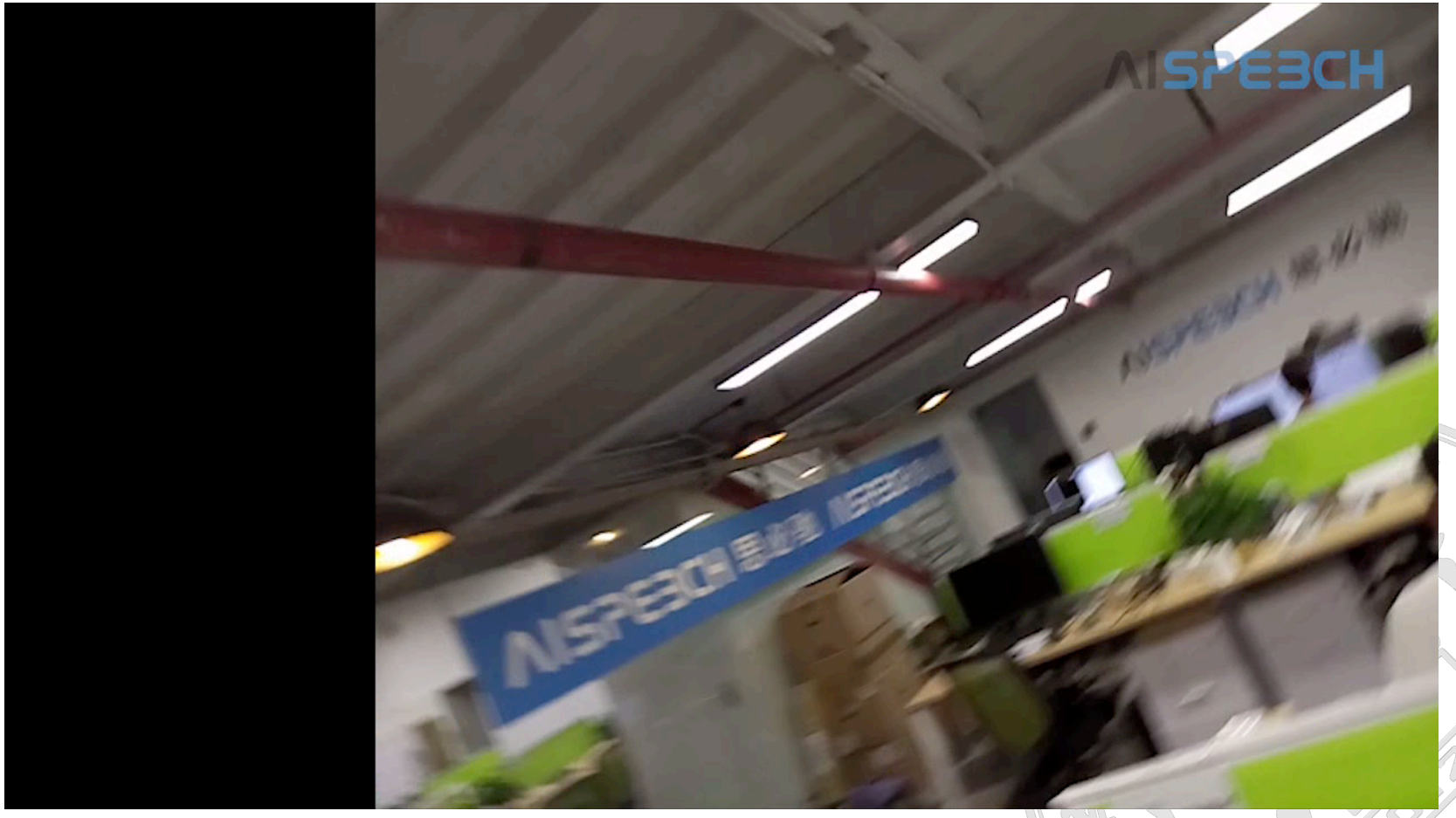
电话语音



任务	识别率（%）	Sample
听写短信输入	>90%	
电话语音	>80%	
网络语音	>70%	

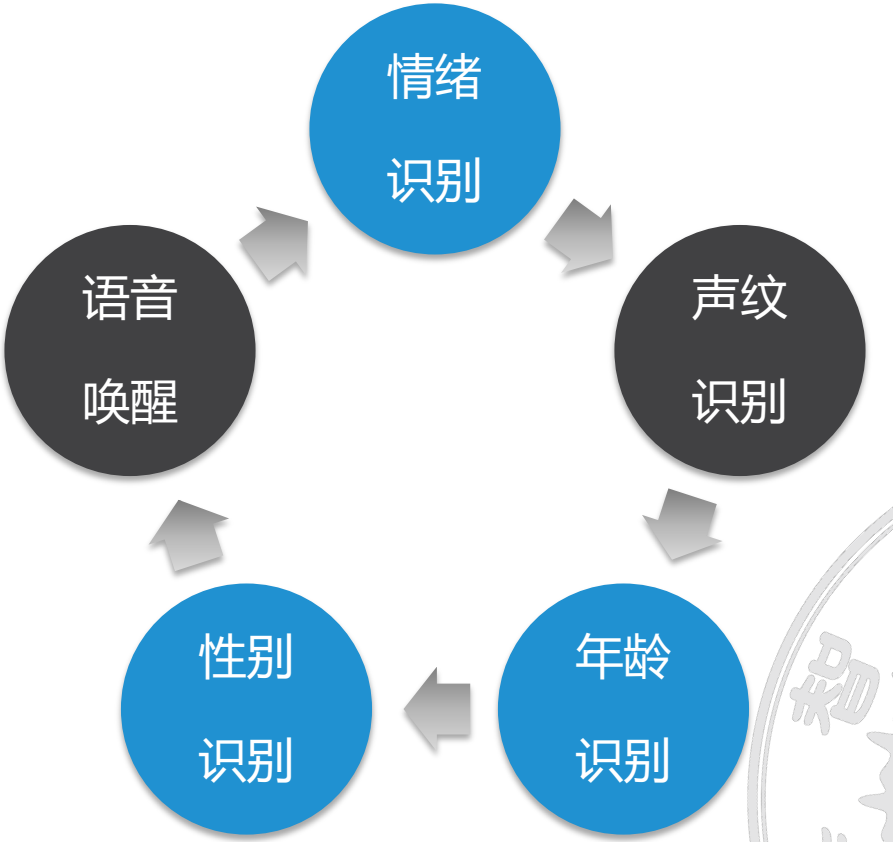
语音识别错误率有多高？

远场噪声环境



弦外之音 —— 丰富音频分析


语音密码：你的声音就是一把钥匙
注册
登录
输入用户名
密码录音时请说
“我是xxx”
两次录音内容需一致
初始化成功


情绪
识别
声纹
识别
年龄
识别
性别
识别
语音
唤醒


语音密码：你的声音就是一把钥匙

声纹识别

输入——

- 音频:
  - 一个整数或实数序列
  - 16000 次采样/秒 (16K Hz)
  - $\boldsymbol{x} = \{x_1, \ldots, x_T\}$
- 特征提取:
  - 25ms 窗宽
  - 10ms 帧移
  - 一个**实数向量**序列
  - $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$

输出——

- 识别：说话人ID
- 验证：是/否

难点

- 变长序列（音频长短差异大）
- 语言及环境干扰

机器学习类型

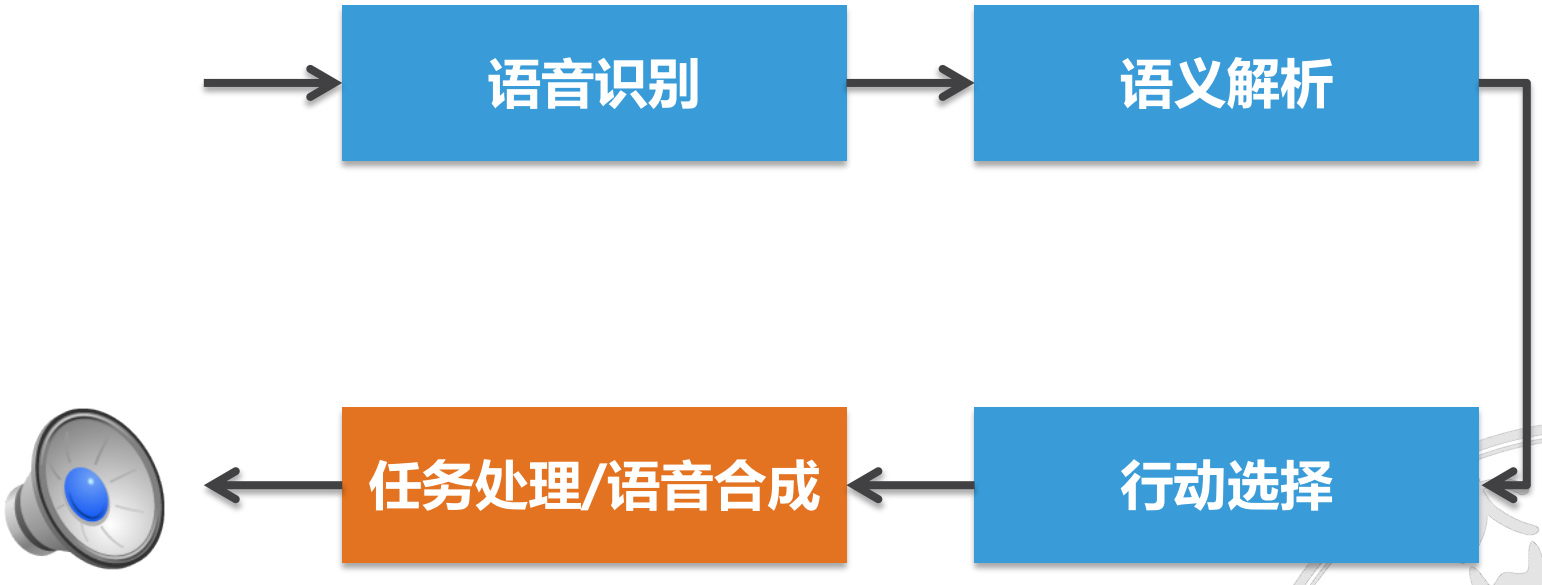
- 验证：分类
- 识别：分类（带拒识/多标签）

声纹与识别，哪个更难？

声纹识别

多人单通道声纹识别

经典人机口语对话系统架构



观点：**合成**就是传递语义信息，只需要**清晰准确**

语音合成

输入——

- 文字序列:
  - 一个带标记的词或音素序列

输出——

- 频谱特征:
  - 25ms 窗宽
  - 5ms 帧移
  - 一个**实数向量**序列
  - $\boldsymbol{O} = \{\boldsymbol{o}_1, \dots, \boldsymbol{o}_T\}$
- 音频:
  - 一个整数或实数序列
  - 16000 次采样/秒 (16K Hz)
  - $\boldsymbol{x} = \{x_1, \dots, x_T\}$

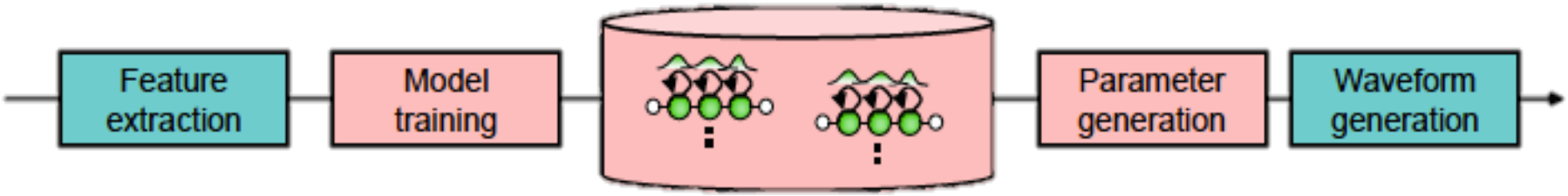
难点

- 变长序列（词序列及音频）
- 副语言信息的丰富和复杂

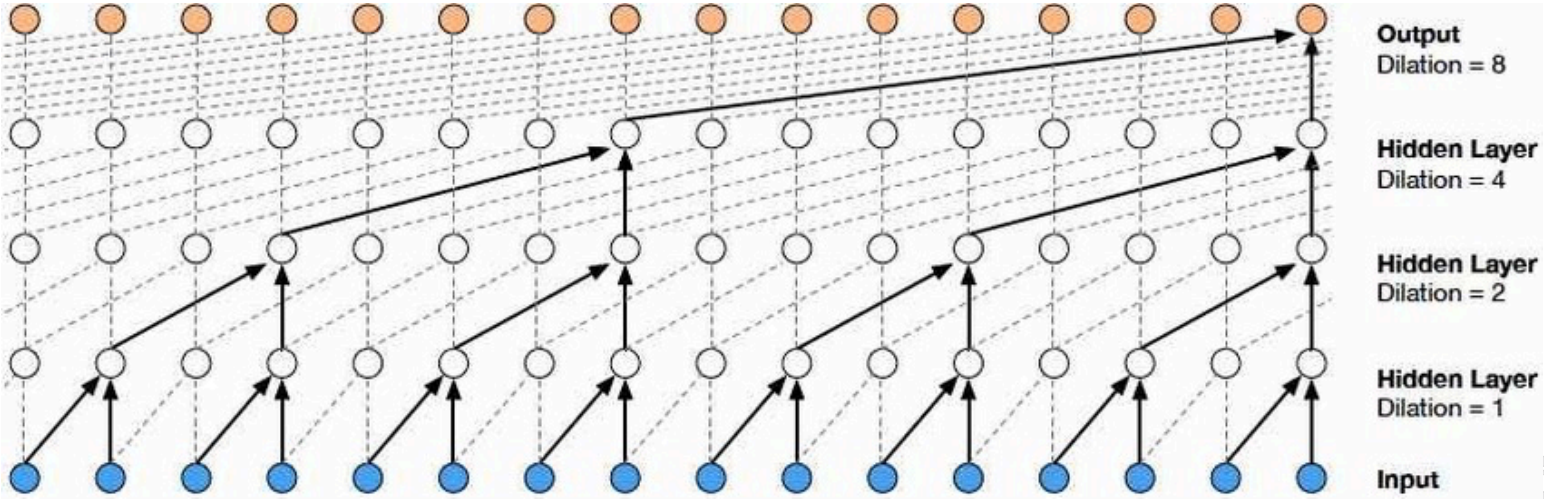
机器学习类型

- 频谱预测：序列生成/回归
- 音频生成：信号处理，序列回归/分类

统计语音合成 —— 数据驱动的个性化语音合成



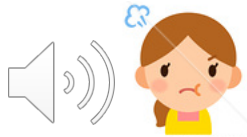
神经网络语音合成 —— 神经声码器与端到端声学模型



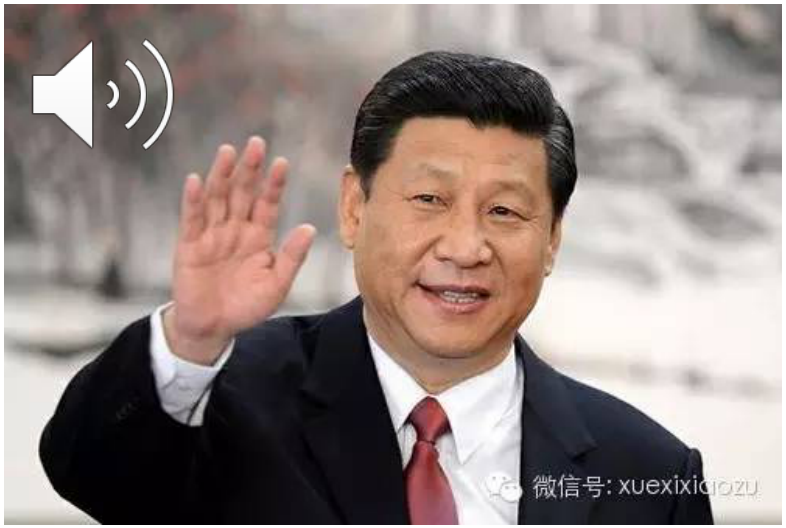
个性化合成 ——语音复刻与操控

Char Sue is an expensive restaurant in the centre.



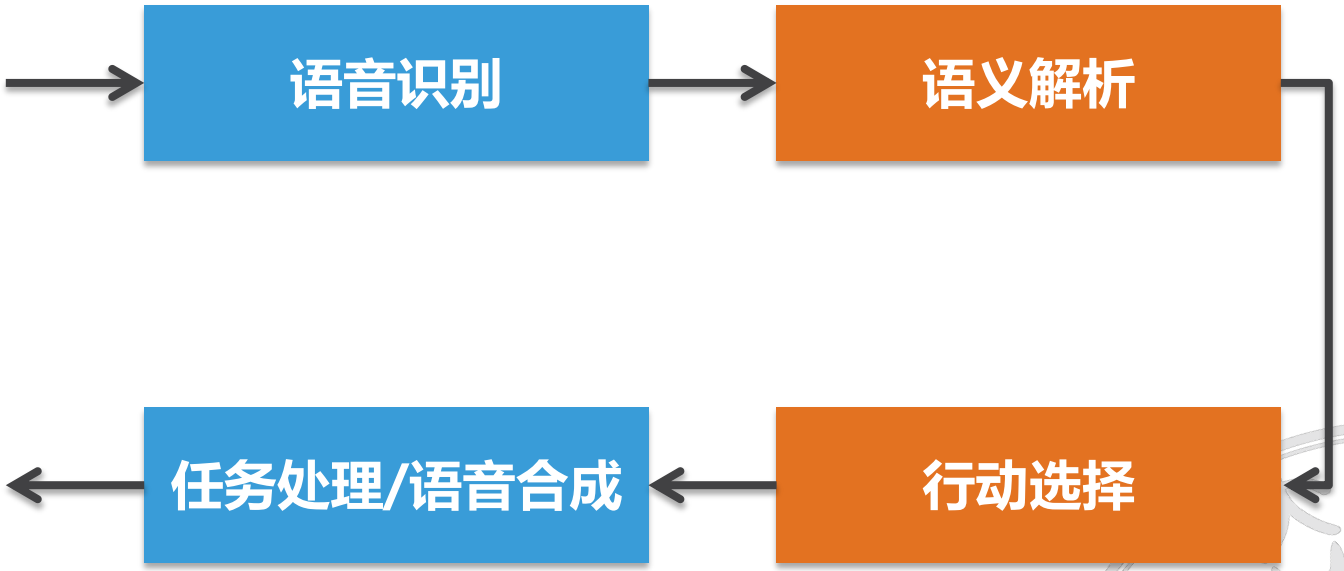


Are the first and last stanzas of Mr. Todhunter's poem the Banshee.



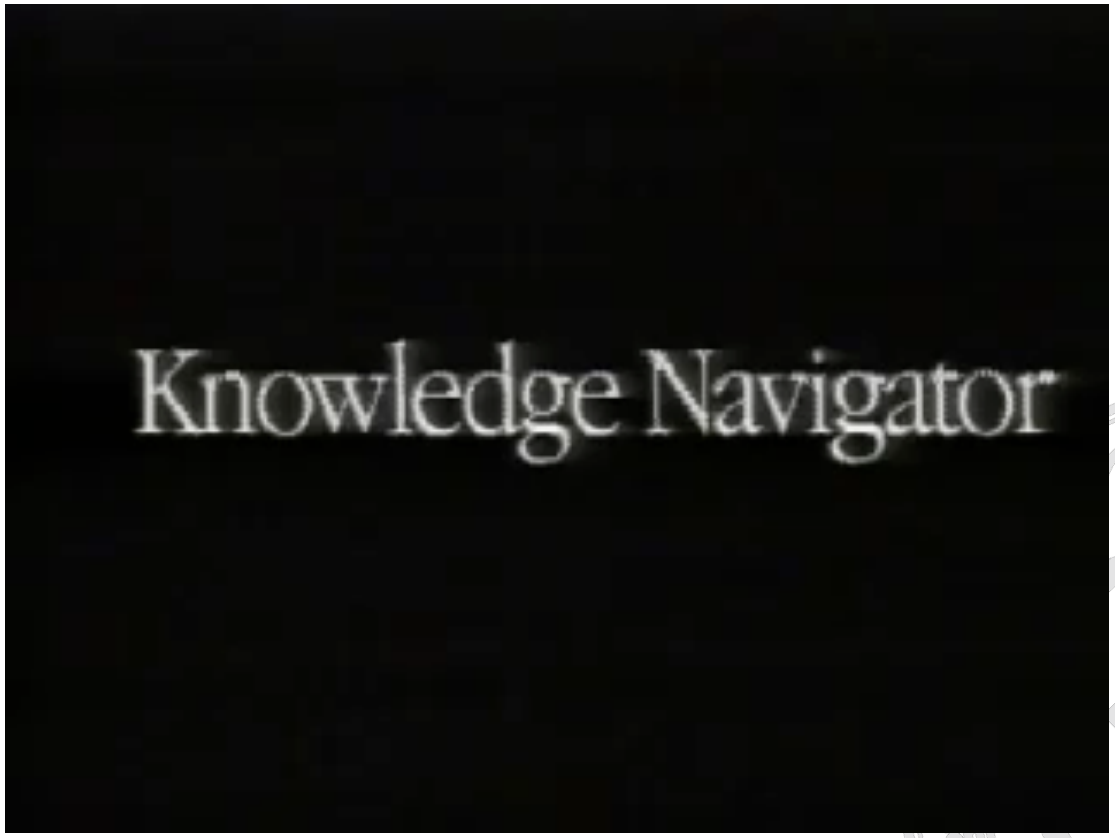
个性化表达 —— 歌曲合成

经典人机口语对话系统架构



观点：**理解**就是文字到语义项的**映射**，**决策**基于预定义**规则**

回到未来



未来已来吗？

口语对话系统的发展路径（美国）

2002 – 2005　EARS

Effective, Affordable Reusable Speech-to-text

2006 – 2011　GALE

Global Autonomous Language Exploitation

2003 – 2008　CALO

Cognitive Assistant that Learns and Organizes

口语对话系统的发展路径（美国）

口语对话系统的发展路径（欧洲）

CLASSiC

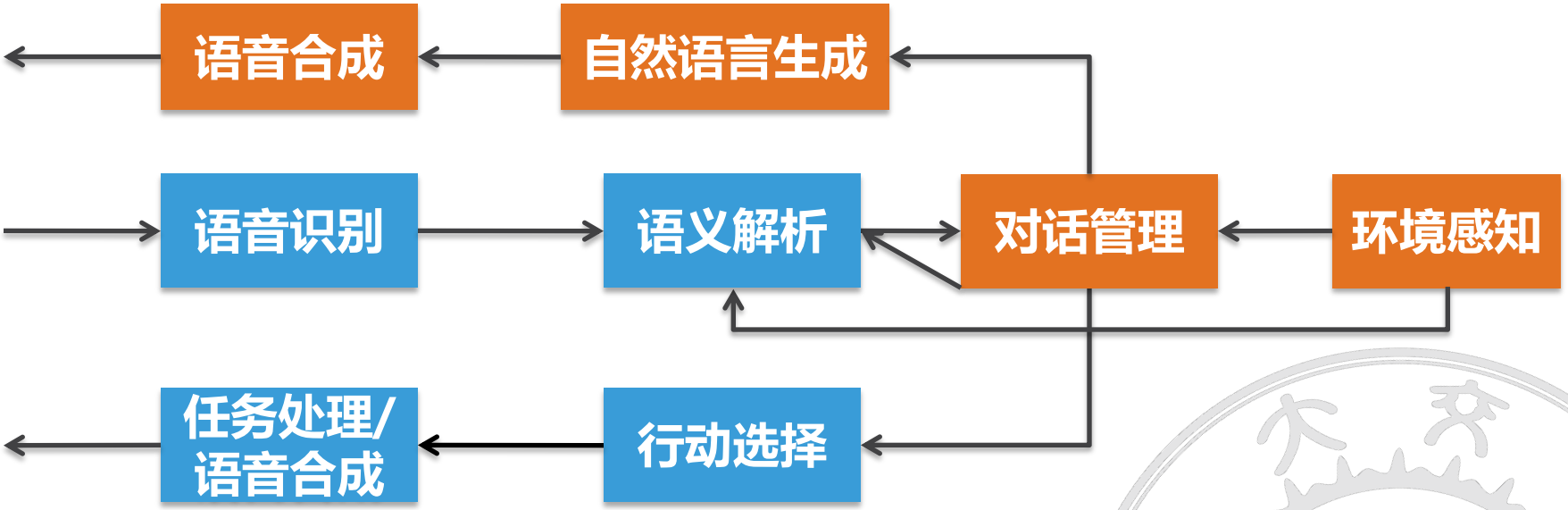
Computational Learning in Adaptive Systems for Spoken Conversation

PARLANCE

Probabilistic Adaptive Real-Time Learning And Natural Conversational Engine

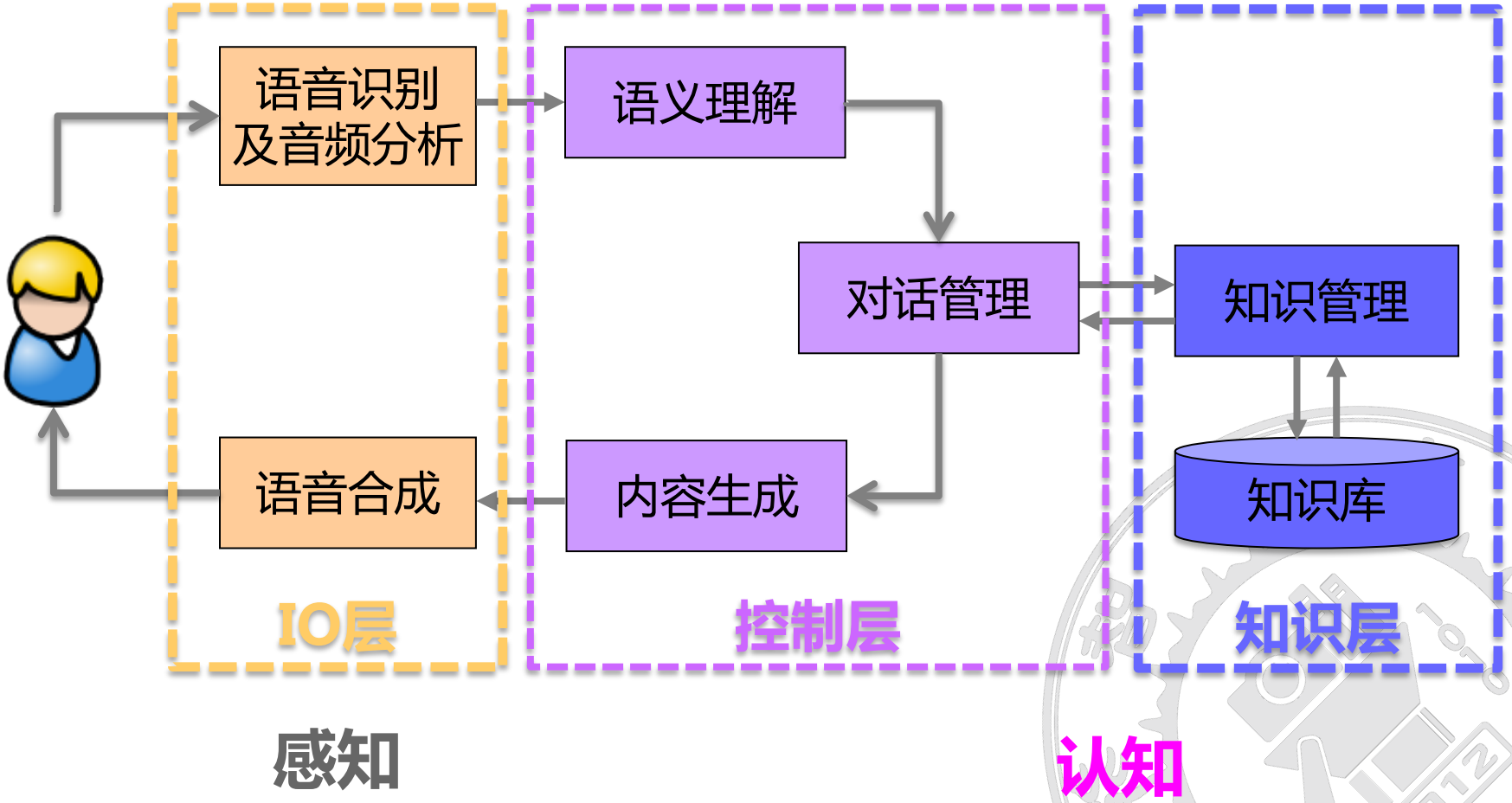
vocal iq

口语对话系统的中枢 —— 对话管理



观点：语音是处理任务的高效管道，不确定性在**多轮交互和情境**中消除

具有认知能力的自然人机口语对话系统



在中国用人工智能改变世界

SJTU Cross Media
Language Intelligence Lab
上海交通大學跨媒體語言智能實驗室

从事智能语音技术的研究，首先要是致力于改变世界的工程师；
而一个杰出的工程师，一定也是一位能够深刻认识世界的科学家。

X-LANCE

从认识世界到改变世界 —— 本课程中将与你为伴的人和事