MDCC 2016
中国移动开发者大会
Mobile Developer Conference China 2016

语音识别现状及有效工具

AISPEECH 思必驰
专注人性化的智能语音

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

钱彦旻

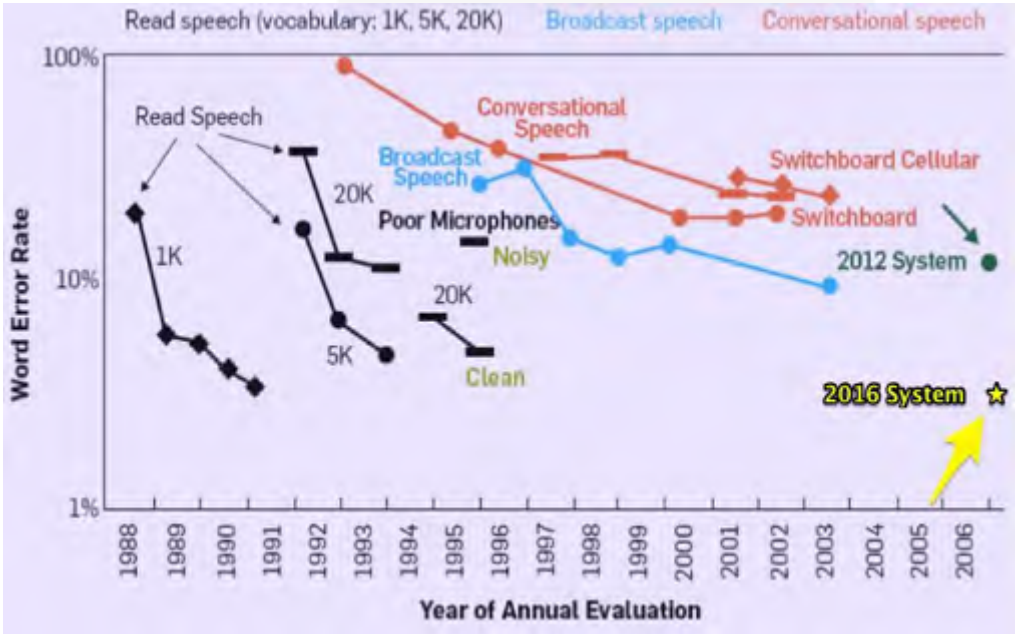
上海交通大学计算机系助理教授
思必驰-上海交大联合实验室副主任

mdcc.csdn.net

CONTENTS

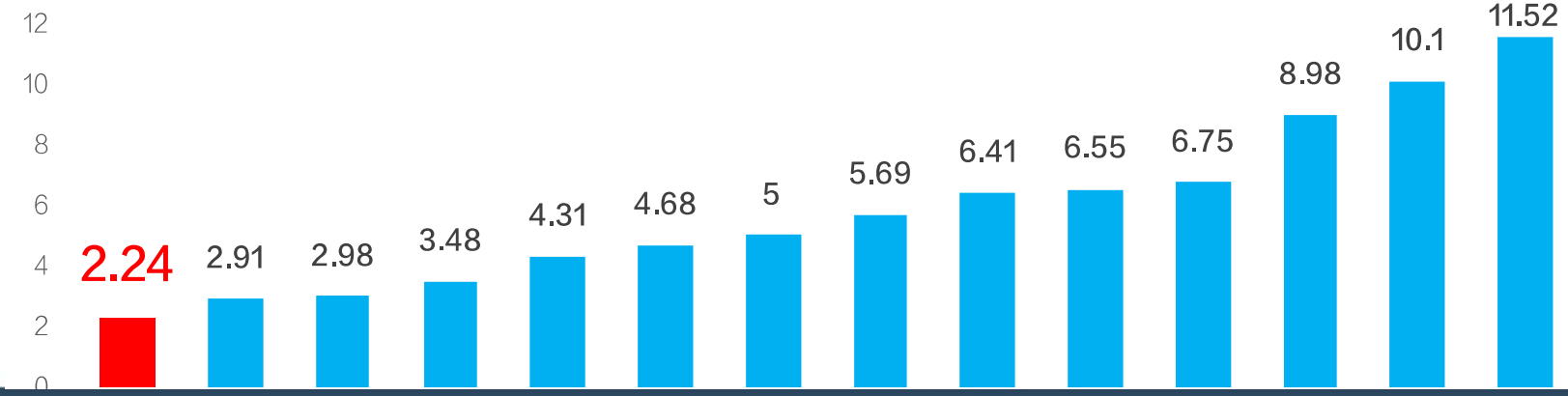
Big News 1 —— last week

- 微软雷德蒙研究院在电话语音识别swbd标准库上到达了 6.3% 错误率
- 人类的能力: ~5.8%错误率



Big News 2 —— last week

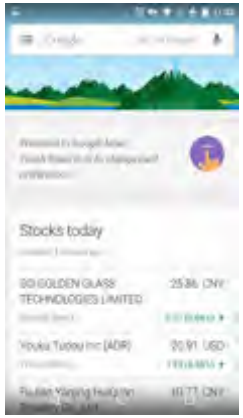
- CHIME-4国际多通道语音分离和识别大赛
- 最好系统性能词错误率已降至~2%



语音识别真的解决了吗？

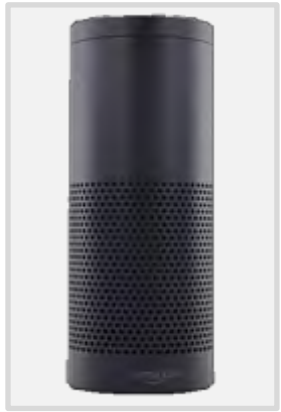
- **Microsoft switchboard system**
  - 电话信道，背景噪声较小
  - Native English Speaker
  - 多遍历的解码策略
  - 多系统后处理融合
- **CHiME-Challenge系统**
  - 朗读语音，小词表，近距离
  - 离线的前端降噪算法
  - 语言模型的过度调优
  - 多系统融合

Siri



Google Now

Cortana


Echo

现实中的语音识别情况

CONTENTS

什么是语音识别

语音识别是把金钥匙-对语音内容进行提取

语音识别的难点

— Variability —

说话人

- Accents
- Dialect
- Style
- Emotion
- Coarticulation
- Reduction
- Pronunciation
- Hesitation
- ……

环 境

- Noise
- Side talk
- Reverberation
- ……

设 备

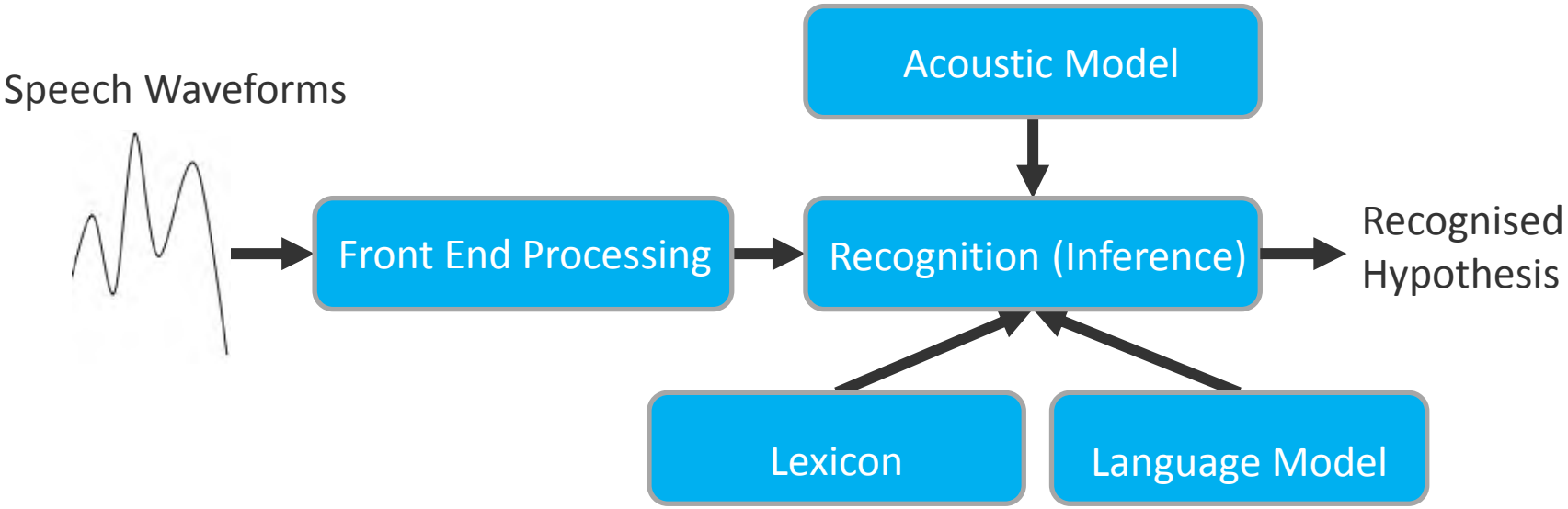
- Head phone
- Land phone
- Speaker phone
- Cell phone
- ……

Interactions between these factors are complicated and nonlinear

统计语音识别

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{O}) = \arg\max_{\mathbf{W}} p(\mathbf{O}|\mathbf{W})P(\mathbf{W})$$

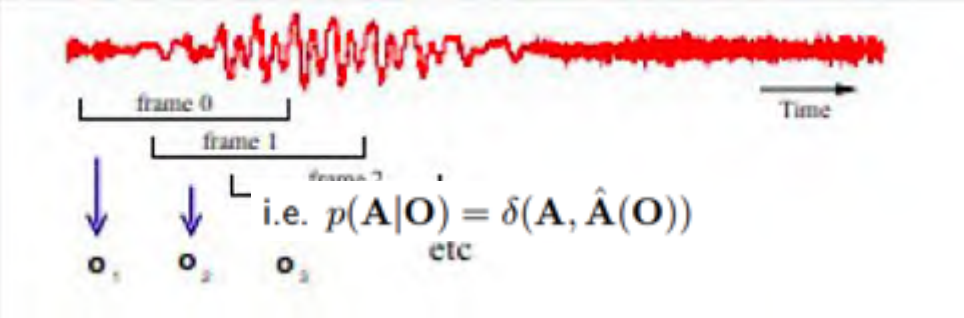
$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{A}|\mathbf{O})p(\mathbf{O}|\mathbf{L})P(\mathbf{L}|\mathbf{W})P(\mathbf{W})$$



特征提取-p(A|O)

原始语音通过信号处理的方法转换成特征向量序列

a sequence of feature vectors.



- 特征提取是一个确定的过程， i.e. $p(\mathbf{A}|\mathbf{O}) = \delta(\mathbf{A}, \hat{\mathbf{A}}(\mathbf{O}))$
- 降低信息率，但是保留有用信息
- 去除噪声或者其他的无关信息
  - 识别原因：最低的两个共振峰
  - 识别性别：隐掉（pitch）或者基音周期频率

声学模型-p(O|L)

声学模型是一个概率模型，它可以描述不通声音的各种不同特性。

- 语音识别是最关键的技术之一。
- 概率模型p(O|L)用户刻画不同语音单元，如音素、音节、字、词。
- Hidden Markov Model(HMM)隐含马尔科夫模型，被广泛采用。

HMM被认为是一个最基本的有限状态传输机，可以将一个用于表示语音的特征向量序列，通过有限状态机，转换成状态机的状态序列，包括音素、音节、词。

字典模型-p(L|W)

字典模型为声学模型和语言模型之间构建了桥梁。

- 它在词和声学单元之间定义了一个映射。
- 它可以是一个确定化的模型（deterministic）。

Word	Pronunciation
TOMATO	t ah m aa t ow t ah m ey t ow
COVERAGE	k ah v er ah jh k ah v r ah jh

- 它可以是一个概率模型（probabilistic）。

Word	Pronunciation	Probability
TOMATO	t ah m aa t ow t ah m ey t ow	0.45 0.55
COVERAGE	k ah v er ah jh k ah v r ah jh	0.65 0.35

语言模型-p(W)

语言模型是一个概率模型probabilistic model:

- 引导搜索算法（在给定历史的情况下预测下一个词）。
- 消除声学单元之间的混淆性，特别是那些声学层相似的单元。

Great wine vs Grey twine

语言模型将概率加到词序列串上去：

- 上下文自由语法

（<s> <one | two | three> </s>）

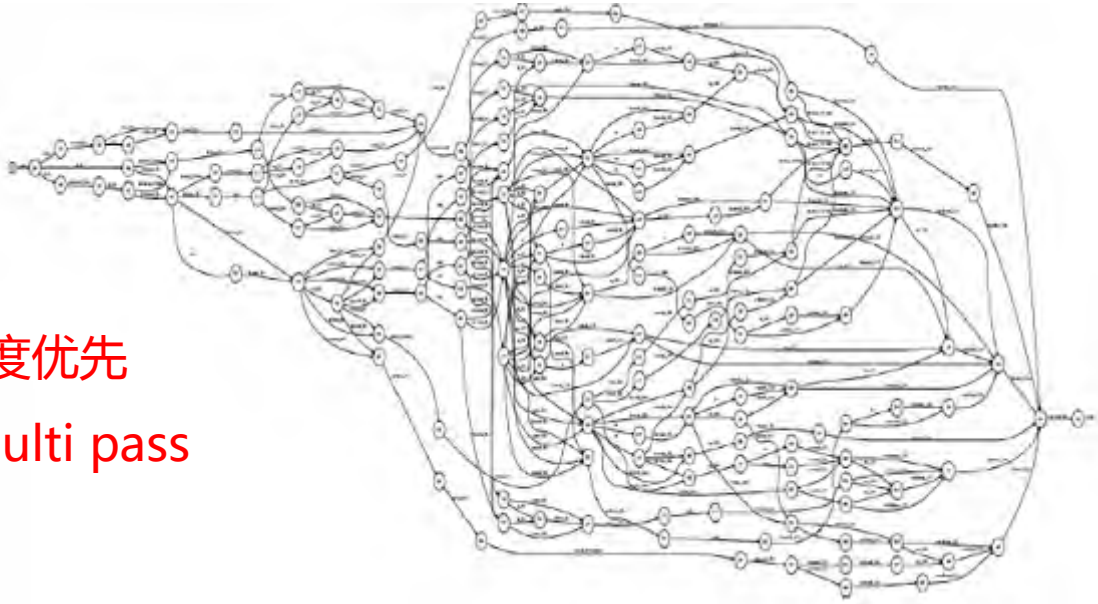
- 统计语言模型：n-gram 语言模型

$$P(w_1, w_2, \ldots, w_N)$$

n-gram统计语言模型和HMM声学模型被广泛运用于语音识别中。

解码和搜索

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{A}|\mathbf{O})p(\mathbf{O}|\mathbf{L})P(\mathbf{L}|\mathbf{W})P(\mathbf{W})$$



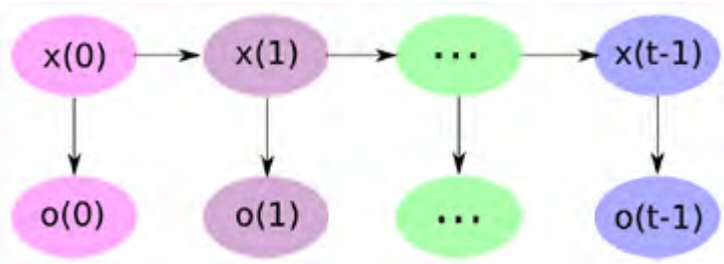
解码算法

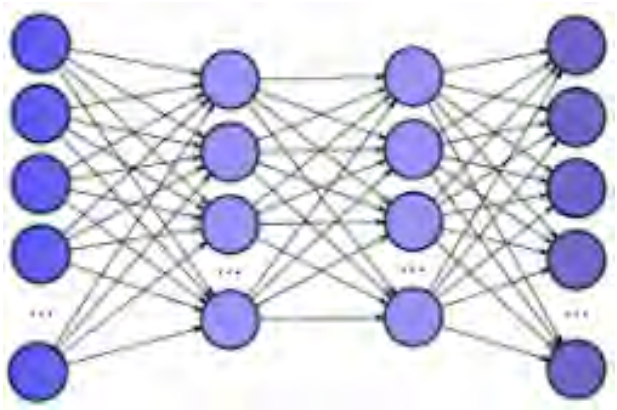
- 动态 vs. 静态
- 深度优先 vs. 广度优先
- One pass vs. Multi pass

传统语音识别

基于深度学习的语音识别

DSP → DNN → Language Model



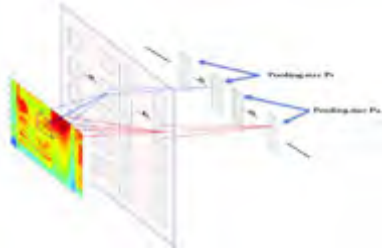


从浅到深：MS, Google, IBM

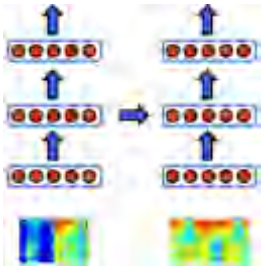
Task	training data (h)	DNN-HMM (%)	GMM-HMM (%)
Switchboard(test set 1) Switchboard(test set 2)	309 309	18.5 16.1	27.4 23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5870	12.3	---
Youtube	1400	47.6	52.3

更强大的神经网络结构

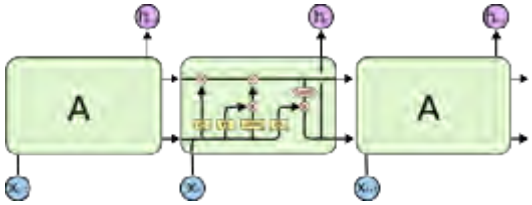
卷积神经网络（CNN）



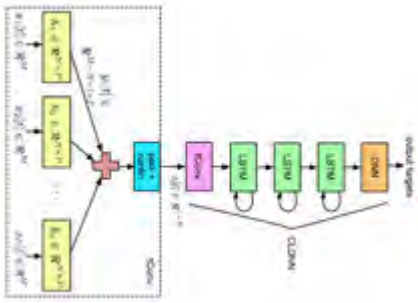
循环神经网络（RNN）

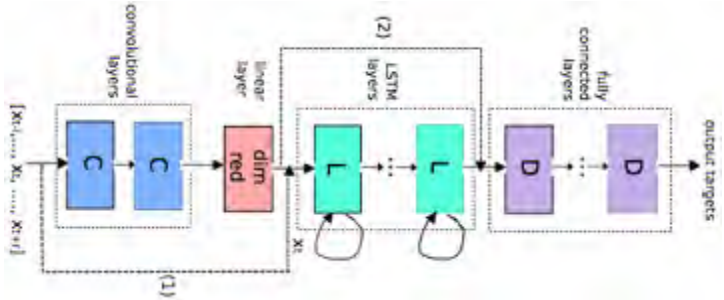


长短时记忆网络（LSTM）

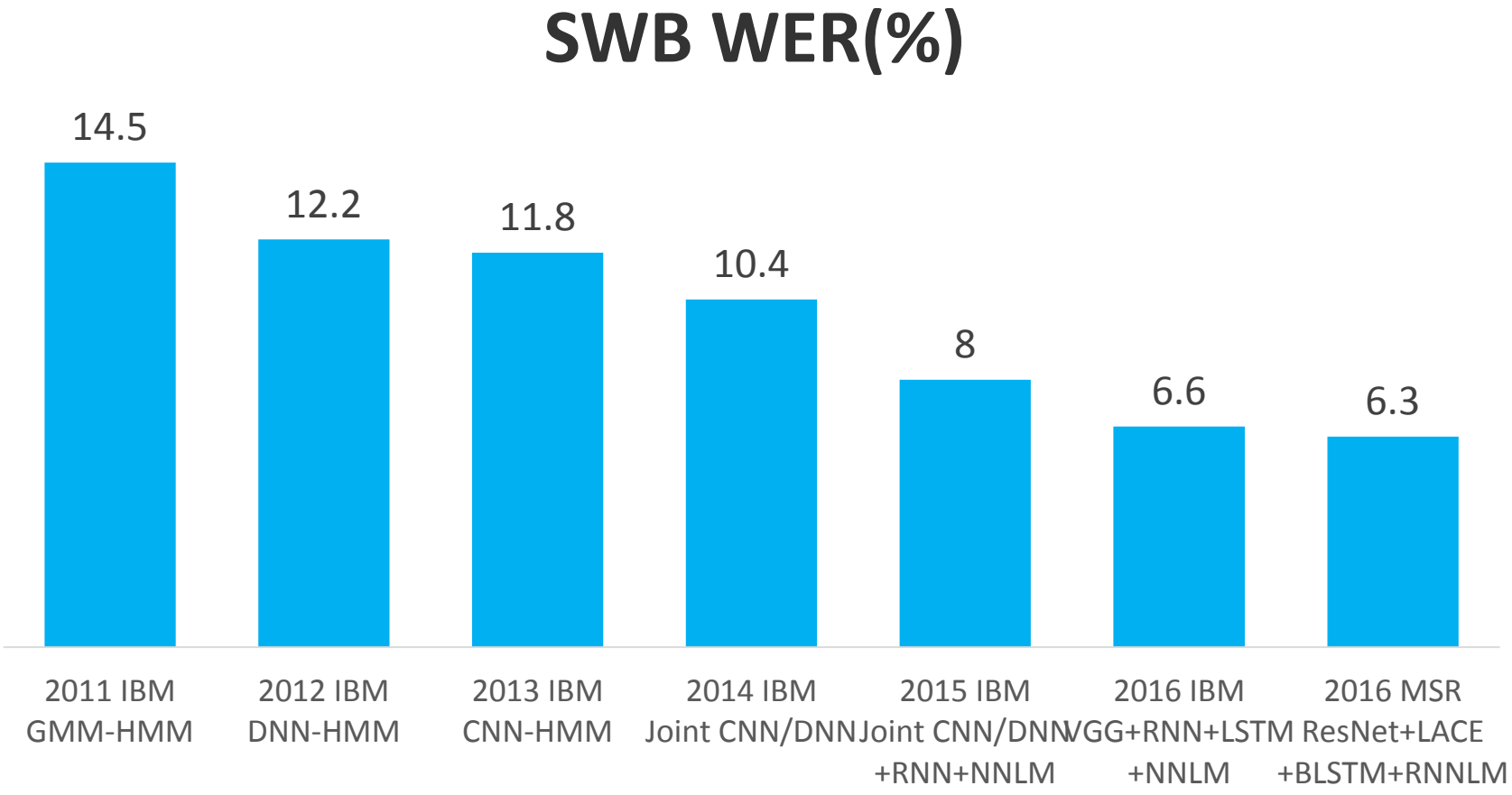


各种网络结构的组合



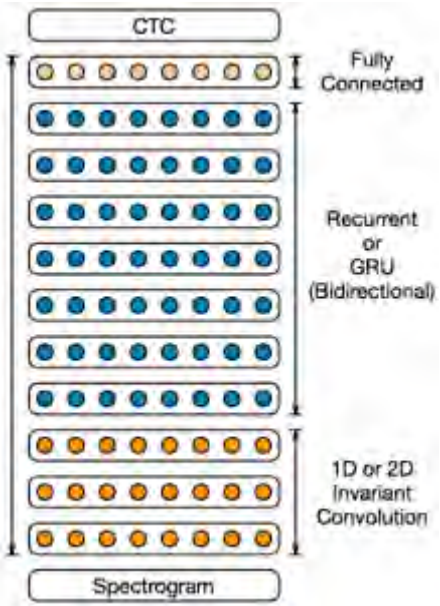


Switchboard电话语音识别发展历程



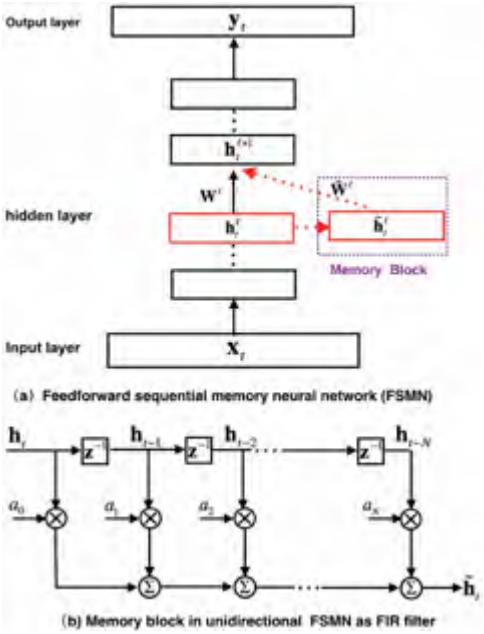
国内同行-公开发表文献可查

Baidu



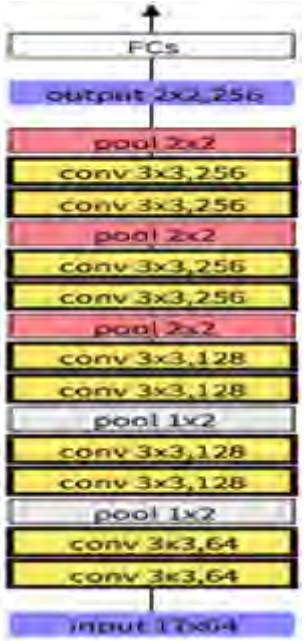
CLDNN

USTC-iFLYTEC



FSMNN

SJTU-AISpeech



VDCNN

语音识别仍面临很多困境

- 噪声鲁棒性
- 多类复杂性
- 低数据资源
- 多语言特性
- 低计算资源
- ……

噪声鲁棒性

噪声环境下的鲁棒语音识别——

- 大规模应用主要绊脚石
- 噪声，远场，信道失配...

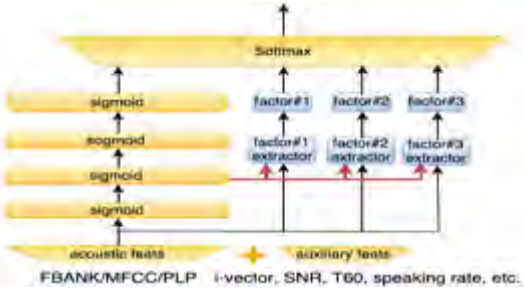


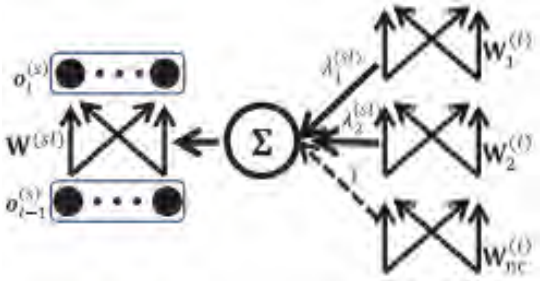


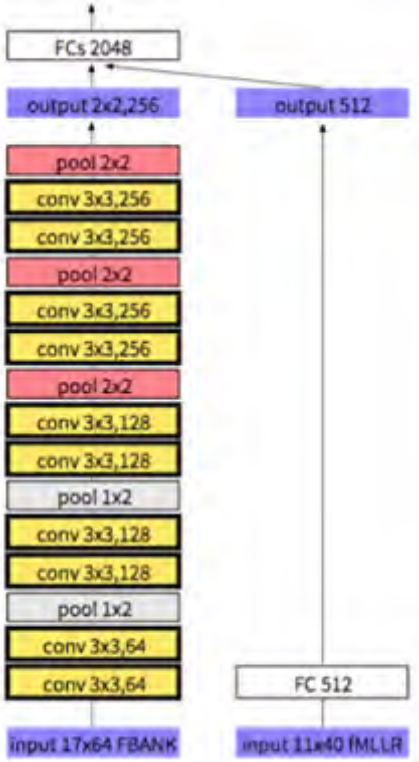
噪声鲁棒性

无论强噪声还是远场都取得了优异性能

- 环境感知深度模型
- 神经网络快速自适应
- 深层卷积神经网络







Cluster Adaptive Training for Deep Neural Network Based Acoustic Model

Tian Tan, *Student Member, IEEE*, Yanmin Qian, *Member, IEEE*, and Kai Yu, *Senior Member, IEEE*

Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition

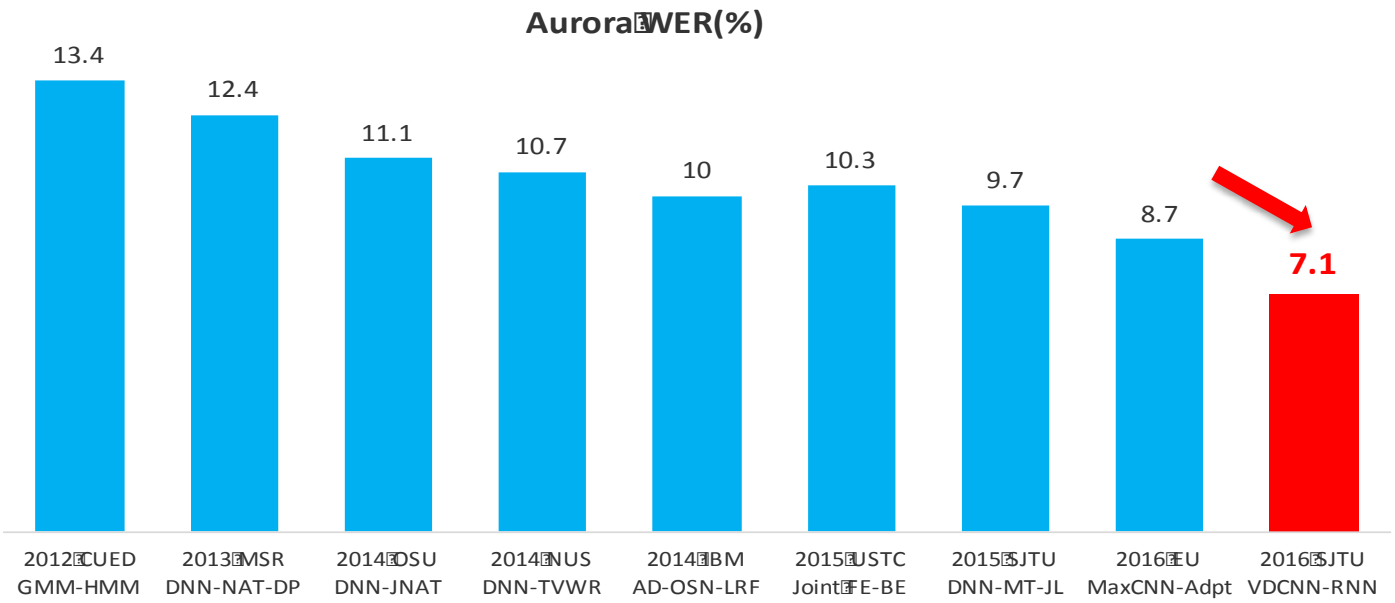
Yanmin Qian, *Member, IEEE*, Mengxiao Bi, *Student Member, IEEE*, Tian Tan, *Student Member, IEEE*, and Kai Yu, *Senior Member, IEEE*

Neural Network Based Multi-Factor Aware Joint Training for Robust Speech Recognition

Yanmin Qian, *Member, IEEE*, Tian Tan, *Student Member, IEEE*, and Dong Yu, *Senior Member, IEEE*

噪声鲁棒性

Aurora4



多类复杂性

多类别复杂语境下的语音识别系统——

Youtube, BBC, etc

多类复杂性

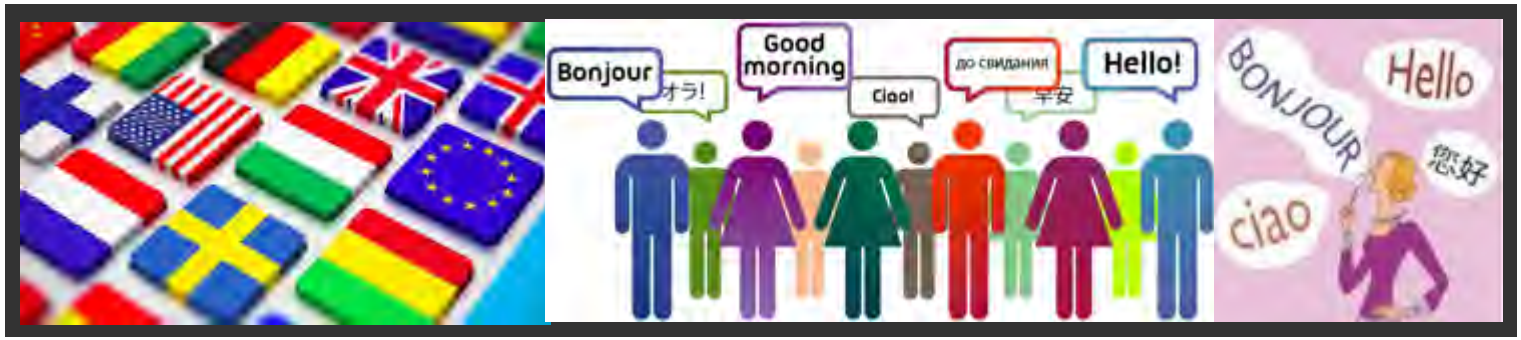
Multi-Genre Broadcast Data Recognition Challenge

- 2015年BBC和EPSRC组办的国际比赛
- 4个单项均列世界第一，且均大幅领先第二名

- **语音识别**
- **说话人分割聚类**
- **标注对齐**
- **时序渐进语音识别**

低数据资源与多语言

多语言及低数据资源小语种语音识别

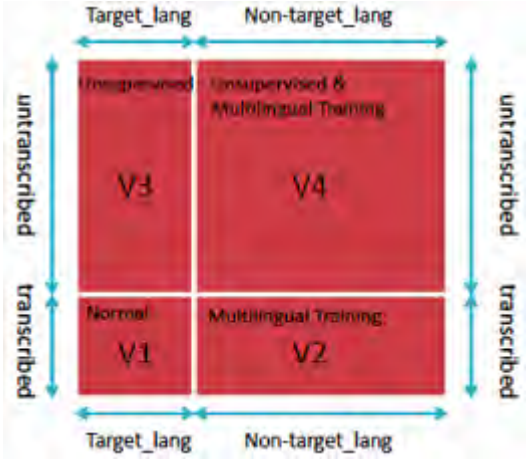


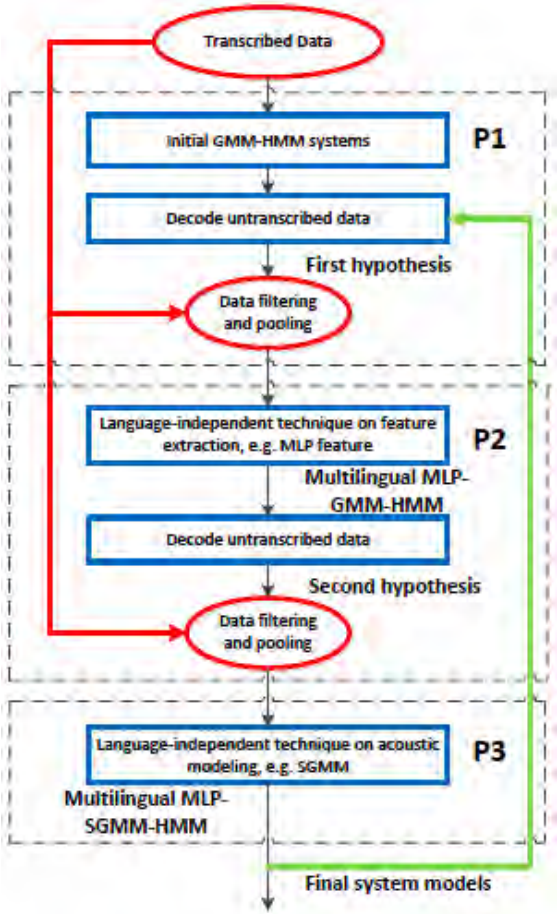
具有战略意义的重要研究方向

- 美国IARPA的Bable计划，美国DARPA的GALE计划
- 中国国情，多民族的特性：56个民族

低数据资源与多语言

各文献报道	词错误率（%）
美国约翰霍普金斯大学JHU	62.8
上海交通大学SJTU	56.5





低计算资源

语音芯片：硬件-软件-服务综合一体

- 低硬件资源，低功耗，离线，实时
- 定制芯片，低成本
- 大词汇量连续语音识别
- 说话人、声纹识别、语音情感

低计算资源

- 连接时序模型取代隐马尔科夫模型
- 音素同步解码取代帧同步解码

model	search step	CER	RTF
HMM	frame	13.3	0.32
CTC	frame	10.2	0.044**(7.3X)**
	phone	10.1	0.016**(20X)**

Phone Synchronous Decoding with CTC Lattice

Zhehuai Chen[1], *Wei Deng*[2], *Tao Xu*[2], *Kai Yu*[1]

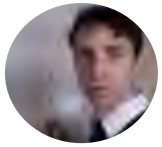
[1]Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]AISpeech Ltd.

{zhenzhehuai,kai.yu}@sjtu.edu.cn, {wei.deng,tao.xu}@aispeech.com

Phone Synchronous Speech Recognition with CTC Lattice

Zhehuai Chen, *Student Member, IEEE*, Yimeng Zhuang, Yanmin Qian, *Member, IEEE*, and Kai Yu, *Senior Member, IEEE*

CONTENTS

KALDI 

- 9 家著名语音研究机构（微软，IBM，IDIAP，SRI，CRIM，布尔诺理工大学BUT，爱丁堡大学，卡尔斯鲁厄大学，清华大学(上海交大)）
- 13 人核心国际工作组（美国、德国、瑞士、英国、加拿大、捷克、中国）
- 2011发布以来，下载量已超20,000,合著的论文已被引用1000多次



Kaldi的特点与影响

【特点】

- 第一个完全用C++编写的语音识别开源工具包
- 第一个完全基于加权有限状态机理论的语音识别开源工具包
- 模块化与高度可扩展性设计,详细的说明文档,完备公开的实验例程
- Kaldi=HTK+SRILM+QUICKNET+RNNLM+HTS……

【影响】

- 被业界广泛采用的标准工具, Apache 2.0，
- 学术界：MIT, CMU,JHU, Cambridge, THU, SJTU等
- 工业界：MS, Google, IBM, Facebook等
- 极大推进了整个语音识别领域的发展

HTK-Hidden Markov Model Toolkit

在剑桥大学开发，第一个语音识别开源工具

- Speech recognition & speech synthesis
- ANSI C,400多页的文档
- 10万多注册用户，5000多次引用

htk3

UNIVERSITY OF CAMBRIDGE

- 历史 (1989-)
  - 1995,V1.5:HMM
  - 1999,V2.2:MLLR,MAP
  - 2000,V3:VTLN,HLDA
  - 2009,V3.4.1:MPE,Hdecode

- 所构建的系统连续蝉联美国NIST和DARPA评测的冠军
- 统治了语音识别领域将近20年，直到Deep Learning的出现

HTK-V3.5

A General Artificial Neural Network Extension for HTK

C. Zhang & P. C. Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.

{cz277,pcw}@eng.cam.ac.uk

HTKV3.5-2015年发布

- 通用神经网络结构的支持
- 基于神经网络的自适应技术
- 基于神经网络的鉴别性训练技术
- Release soon: CNN,GRU,LSTM

使用HTK-V3.5所构建的系统获得了多个世界性评测的冠军

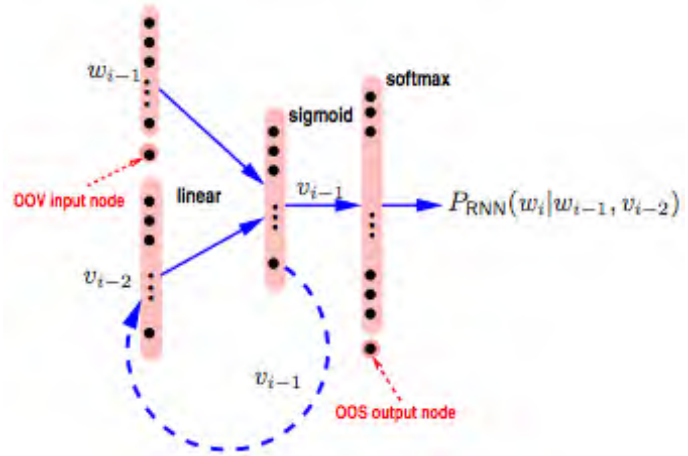
- 2014：DARPA-BOLT冠军
- 2014：IARPA-Babel冠军
- 2015：IARPA-Babel冠军
- 2015：EPSRC-MGB冠军
- 2016：IARPA-Babel亚军

CUED-RNNLM

剑桥开发，2015年发布

- CUDA并行训练方案
  - Class/Full output
  - Minibatch training with GPU
- 快速训练和评估准则
  - Standard CE / VR / NCE
- RNNLM自适应技术
- RNNLM与HTK3.5&Kaldi的结合
  - Lattice rescoring
  - Support HTK lattice directly
  - Support Kaldi lattice
- 详细的文档和完整Recipe
- 用于剑桥近期的各个比赛系统

UNIVERSITY OF CAMBRIDGE



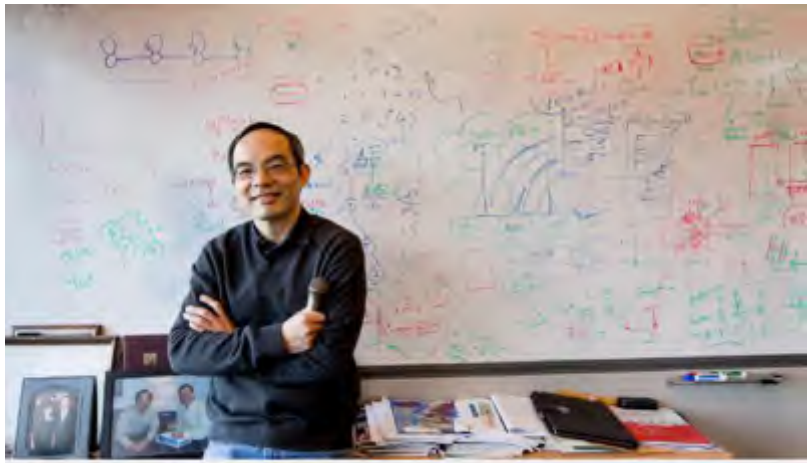
CUED-RNNLM – AN OPEN-SOURCE TOOLKIT FOR EFFICIENT TRAINING AND EVALUATION OF RECURRENT NEURAL NETWORK LANGUAGE MODELS

X. Chen, X. Liu, Y. Qian, M.J.F. Gales, P.C. Woodland

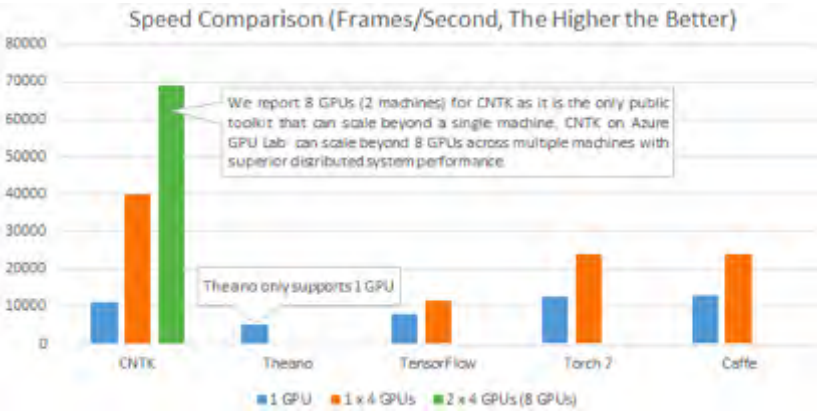
Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ, U.K.

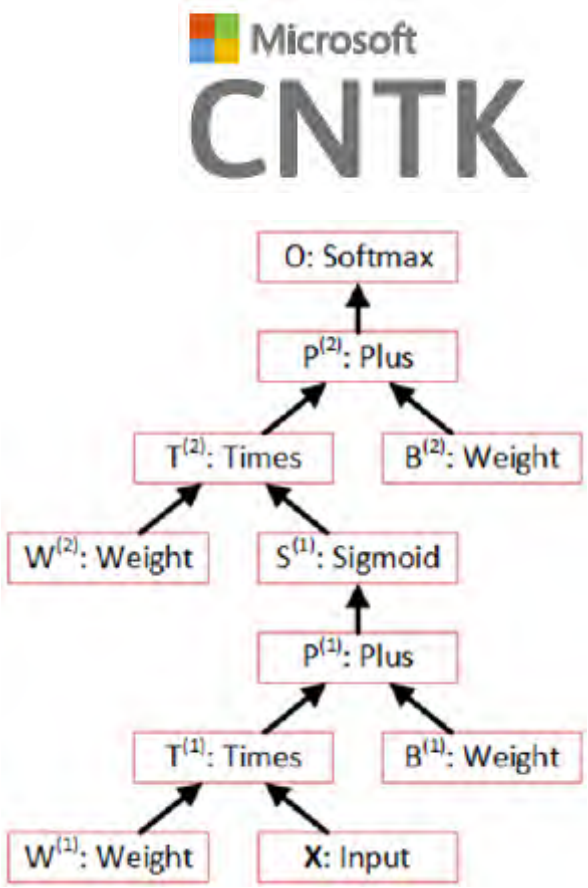
Deep Learning Toolkits

Computational Network Toolkit



Xuedong Huang (Photography by Scott Eklund/Red Box Pictures)

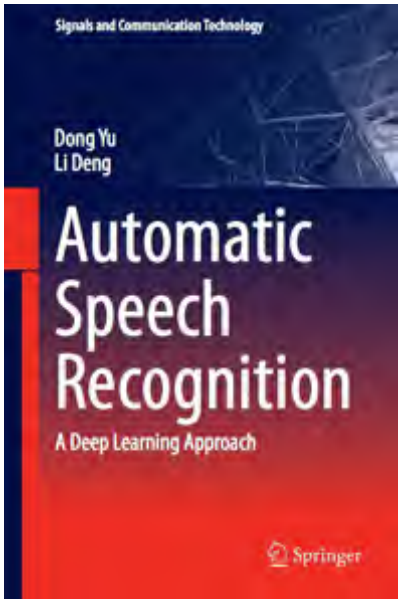




参考书

第一本详细介绍深度学习和语音识别相结合的书籍

2015年英文版出版，2016年中文译版出版



CONTENTS

思必驰：国内极少数的拥有完整自主知识产权的语音公司

国内仅有的两家有全面语音技术公司之一

语音识别

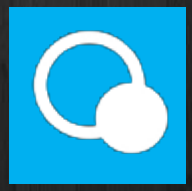
实时云识别
大词汇识别
本地语音识别
抗噪及远场识别

语音合成

童音
名人合成音
标准男女声
个性化定制
高质量合成
超快速合成

语音识别++

语音唤醒
声纹识别
年龄识别
情绪识别



语义理解

电话短信
音乐电台
导航周边
天气日历
票务股票
设备控制等

智能对话

自由打断
智能纠错
渐进理解
任务对话
跟踪意图

AISPEECH

思必驰：深耕垂直场景的 语音交互技术

AIOS-人机对话操作系统

语音交互　对话逻辑

内容服务　平台对接

AICHIP-智能语音芯片模组

智能芯片　麦克风阵列

（环形6+1）

（线性4麦）

AISPEECH

智能家居

远场识别 | 算法降噪 | 回声消除 | 声源定位 | 场景对话

AISPEECH

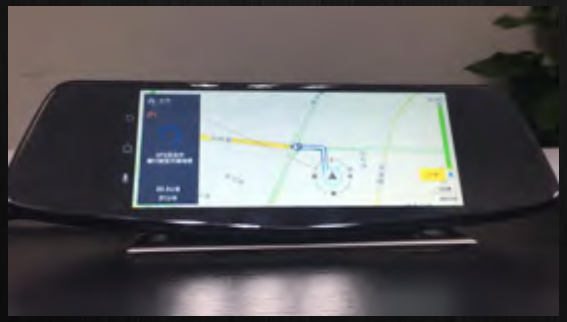
智能机器人

远场交互 | 声源定位 | 回声对消 | 语义对话 | 个性唤醒

AISPEECH

智能车载

云+端交互 | 近场抗噪 | 语音唤醒 | 后端服务 | 场景对话



AISPEECH

AISPEECH

专注体验

整合后端内容/服务，打造更极致的产品体验

技术为本

专注智能硬件，专注自然语言交互

并肩同行

协助整合供应链资源，提供一站式的产业化服务

AISPEECH

QUESTIONS?

THANK YOU

AISPEECH 思必驰
专 注 人 性 化 的 智 能 语 音

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

钱彦旻

上海交通大学计算机系助理教授
思必驰-上海交大联合实验室副主任