互联网时代信息交互与言语信息处理的前沿问题探讨

党建武

天津大学计算机学院

计算机和现代社会

- 第一次工业革命（1780-1800年） 以瓦特的蒸汽机为代表的机械动力革命。
- 第二次工业革命 （1830-1900）以法拉第发现电磁感应为起爆剂的电力技术革命。
- 第三次科技革命是以原子能、电子计算机、空间技术和生物工程的发明和应用为主要标志，涉及信息技术、新能源技术、新材料技术、生物技术、空间技术和海洋技术等诸多领域的一场信息控制技术革命。
- 可以说第三次工业科技革命也是围绕着计算机这个载体来发展的。体力功能的延伸=>器官功能的延伸=>智力功能的延伸。

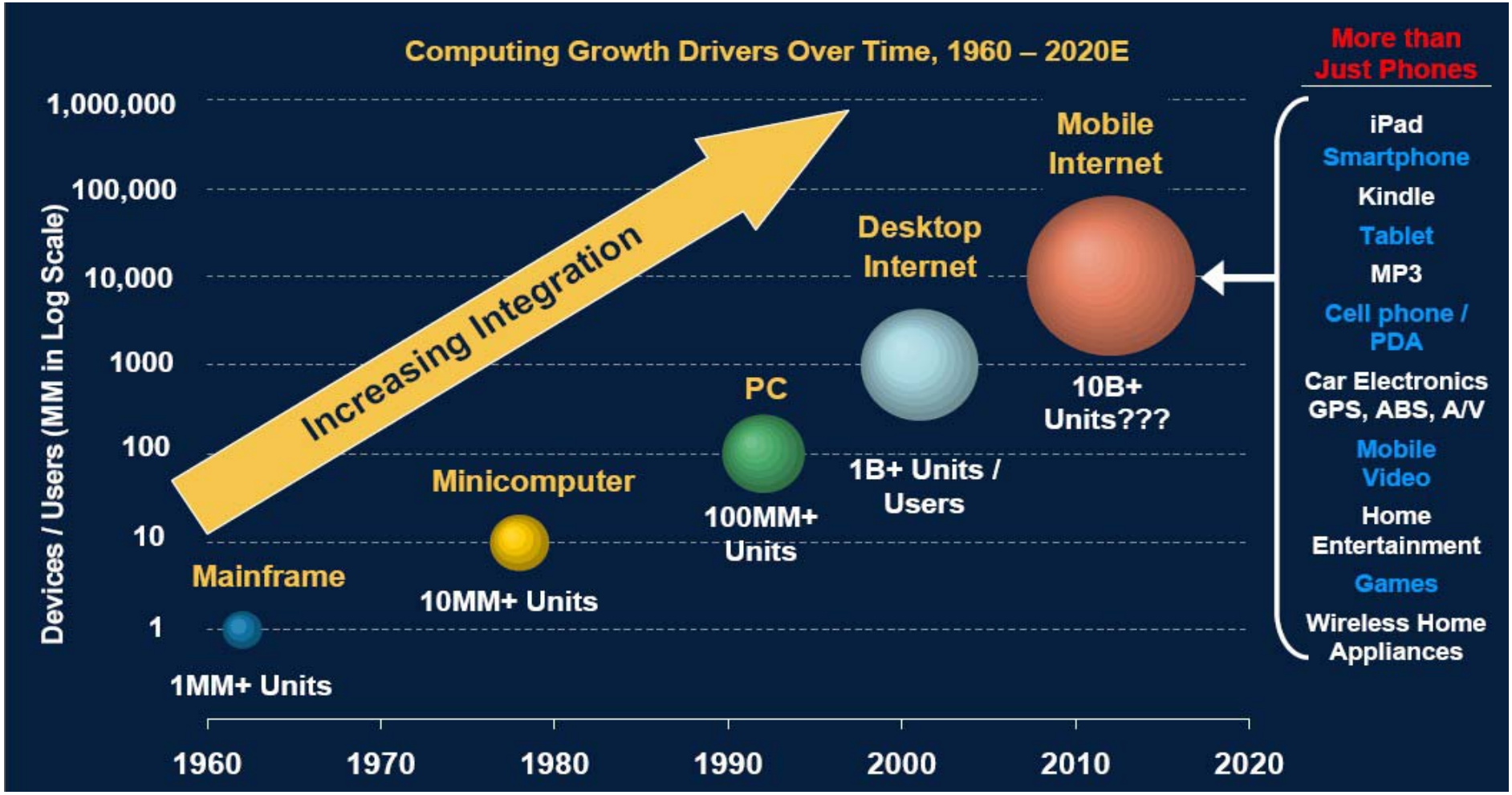
计算机科学的目标

- 计算机的出现使得科学技术从实现人类的技能转变为模拟人类的智能。
- 计算机科学的终极目标，不只是为人类提供一个快捷的计算工具，而是把人从重复、枯燥的信息处理中解脱出来，给人类创造一个具有同样智能的机器伙伴。
- 人类的言语机能是大脑革命的产物，也是人类区别于其他动物的最主要标志之一。语言的出现使得人类将经历积累为经验，将经验升华为知识用语言记录下来，传给下代。
- 描述和实现人类的言语功能和由此而来的知识是计算机科学的重要任务之一。

互联网的发展及作用

- 40年前，计算机网络的产生：数据传播。
- 20年前Web1.0的产生：信息共享。
- 近10年Web2.0的产生：知识共享。
- 获取信息的方式产生了翻天覆地的变化
  - 从最开始数据传递，
  - 上门户网站看新闻，
  - 博客时代，一个个文化名人、草根英雄、码字少年风起云涌，丰富了我们的信息
  - 过渡到了微博时代。交互性改变了知识的产生形式，人人都是知识的贡献者
- 互联网使人们时而生活在现实世界，时而生活在虚拟世界
- 随着互联网的进化，**人们实现了零距离的沟通，但同时因为其虚拟性又感到彼此陌生。**

互联网引发了大数据时代的到来



IBM的研究称，整个人类文明所获得的全部数据中，有90%是过去两年内产生的。而到了2020年，全世界所产生的数据规模将达到今天的44倍。

大数据时代的数据缺失

- 美国：联邦政府（200万名）共拥有848拍字节（10^{15}B）的数据总量。全世界最大的零售巨头沃尔玛，其数据库大小赶不上商务部下属的美国普查局
- 可获得的大数据：来自互联网（民间）
- 来自中国政府部门、机构的大数据？

大数据研发计划

美国政府（DARPA）2012年3月公布了“大数据研发计划”。

目标：提高从海量和复杂的数据中获取知识的能力，加速科学与工程领域发明的步伐，增强国家安全，并改变教学研究。

其中的任务：

- 解决目前情报，监视和侦察系统的不足。
- 实现对自然文本的理解和进行知识标注。
- 如何运用人机交互手段对半结构化和非结构化数据进行自动处理，并可视化。

互联网环境多媒体信息在爆炸性地增长

flickr Picasa 图像

视频 BBC YouTube

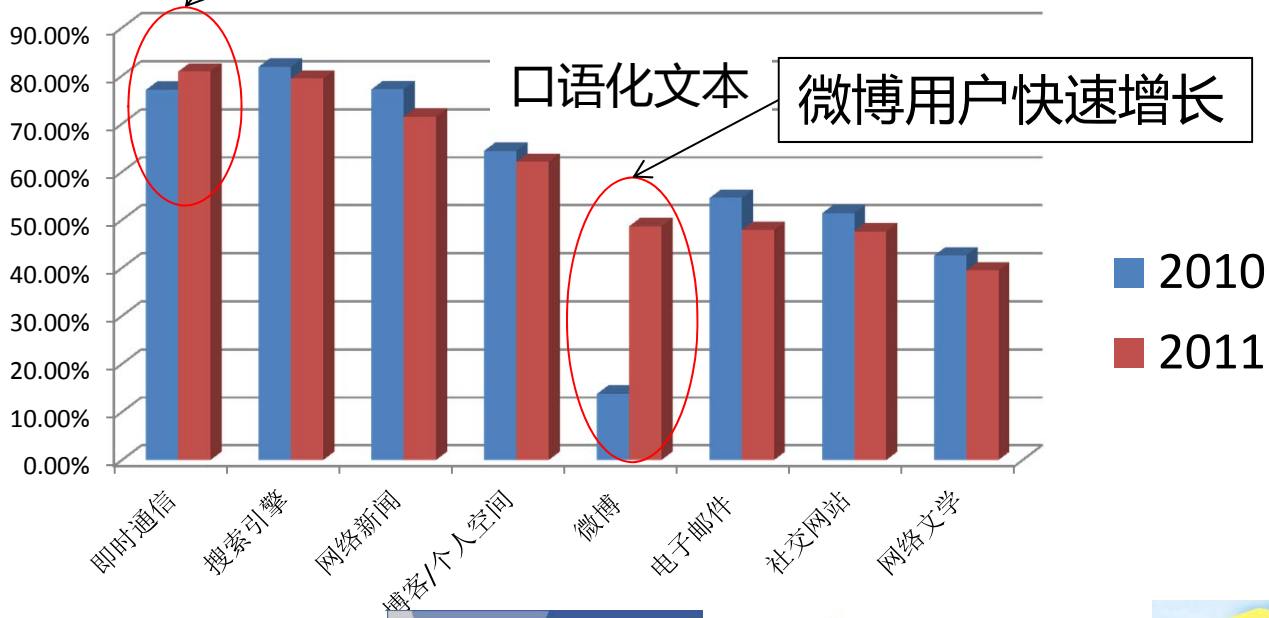
WIKIPEDIA Die freie Enzyklopädie 文本

言语 facebook myspace a place for friends

言语正在成为未来互联网交际的主要媒介

即时通信用户首次成为我国网民的最大群体

语音　我[illegible]12.1）

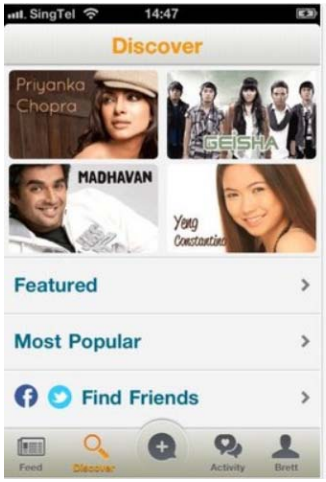


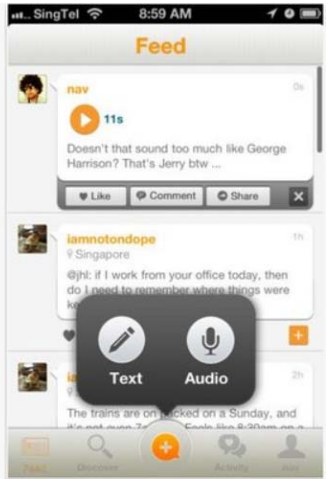
口语化文本

微博用户快速增长

注册用户10亿，同时在线1亿	全球用户突破3亿，活跃用户1.4亿	注册用户9.1亿，活跃用户5.26亿	注册用户突破3亿，活跃用户2700万	手机网民达3.56亿，同比增长17.5%。

互联网言语信息带来的挑战与机遇



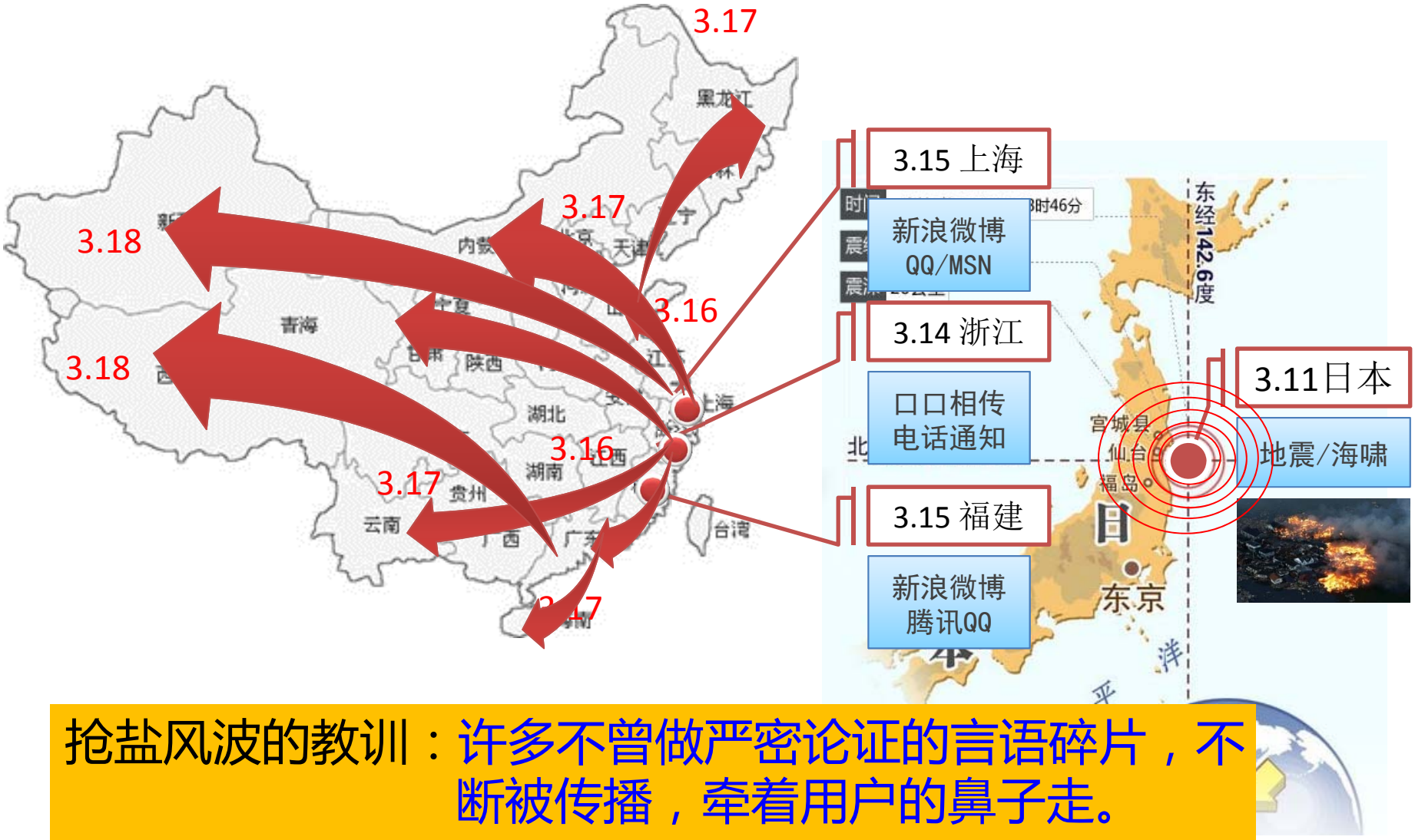


- 微信就是例子
- 支持文本，语音短信，…
- 用户数已经超过一亿，预计2015年将超过三亿

- Bubble Motion是东南亚的一个语音博客平台
- 75%的博客是基于语音的
- 用户数1600多万（2012.3）

言语信息如此迅猛地增长，给国家公共安全带来了挑战，也给信息的深度利用带来了机遇

互联网环境信息的自然传播与影响



抢盐风波的教训：许多不曾做严密论证的言语碎片，不断被传播，牵着用户的鼻子走。

互联网环境信息的自然传播与诱导疏通

3月14日

口口相传

•核辐射传言
•食盐供应不依赖海盐

无需抢盐

自然传播

3月15日

网络言语+即时通讯

•海水污染为谣言
•市场供应充足

无需恐慌

自然传播

3月16日

网络言语+即时通讯

•官方宣布无辐射影响
食盐保证供应

制止蔓延

通过对信息碎片的深度计算有可能及时提供诱导疏通

互联网环境中文言语处理势在必行

消除和缓解
多民族语言间的屏障

研究互联网环境中文言语处理的理论和关键技术势在必行

互联网环境中文信息处理与深度计算的基本理论与方法 (973指南)

- 研究互联网规模中文深度计算的理论与模型，包括中文信息表示理论与模型、句子与篇章的结构分析和语义理解等方法；
- 研究言语多通道感知机理，包括多言语识别、翻译、合成与融合、开放式多类型语言知识大规模获取与组织等方法；
- 研究中文信息所承载中国文化元素的获取、传承和呈现等方法；
- 构建大规模中文文本语义体系和语料库，开发相应软件系统原型。

信息领域支持方向之5
——互联网环境中文信息处理与深度计算的基础理论和方法

互联网环境中文言语信息处理与深度计算的基础理论和方法

党建武 （项目首席科学家）

首席单位：天津大学
参加单位：清华大学　北京大学　北京理工大学
中科院自动化所　中科院软件所
中科院声学所　社科院语言所　西北民族大学

互联网规模：人人参与、人人互动的社会格局（1）

从Web1.0时代发展到Web2.0时代，网络变得具有交互性，形成了社交平台。互联网的社会属性开始清晰地呈现出来。

- 交互的行为主体 用户、开发者、服务方和被服务方之间的边界越来越模糊；原来的信息发布是有中心的，现在人人都是发布者，就没有中心了。互联网的特征之一就是去中心化。
- 交互的表现形态 显性和隐性并存。显性的形态是公开的、*客观的*。而隐性的，我们很难控制虚拟世界带来的正面或负面的影响，利用得好有积极作用，控制不好就会产生极大的负面影响；
- 交互的特点 不像传统的交互，互联网上的交互是丰富多彩的，是碎片化的、随心所欲的，它的范围和影响力就不言而喻了。

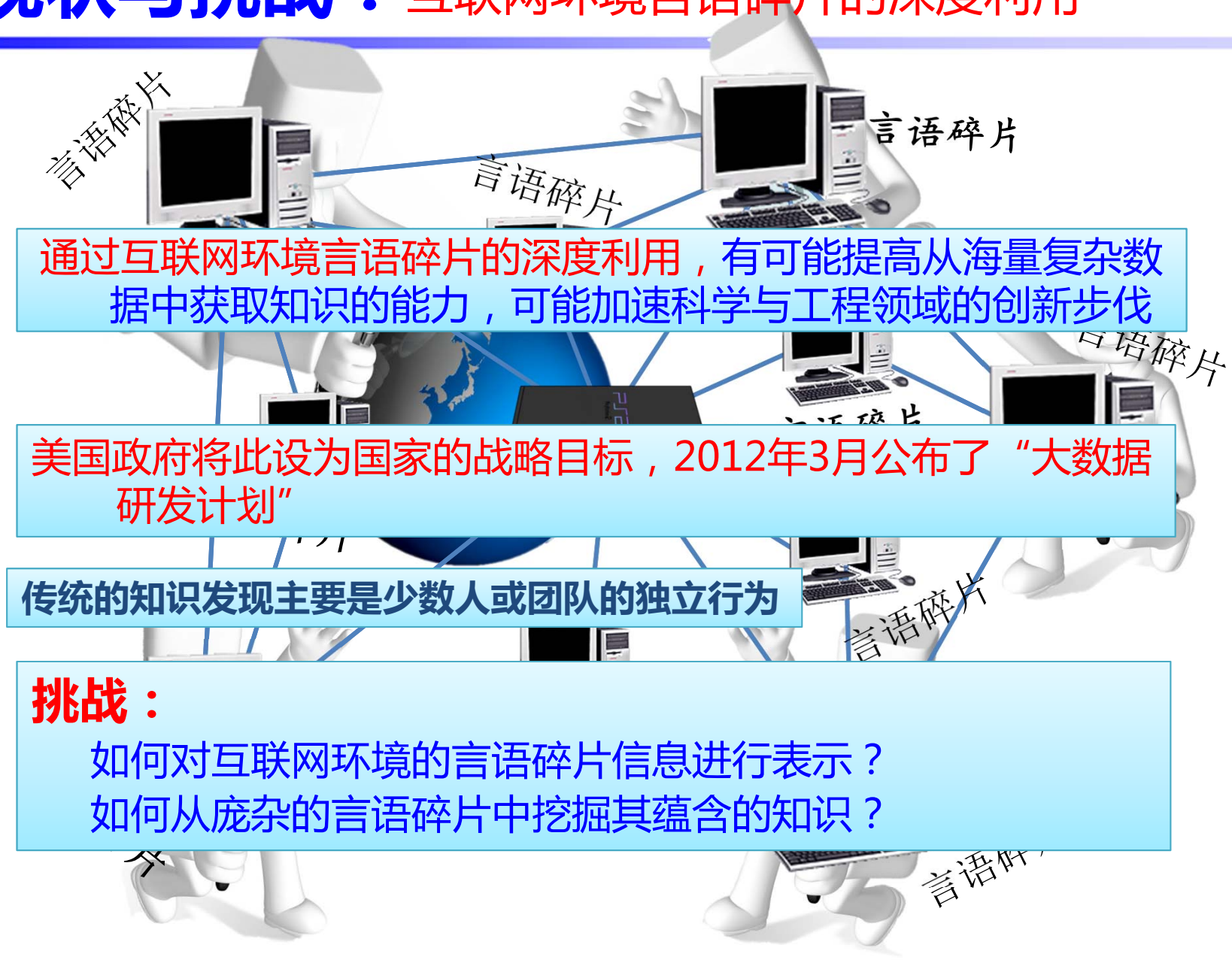
互联网规模：人人参与、人人互动的社会格局（2）

- 交互形态使得社会生产方式有了深刻的变革，例如社交化的教学方式，社交化制造流通方式的变革。
- 互联网技术提供了一个零距离沟通的渠道，提供了新的社交媒体的渠道，形成了新的社会格局。
- 传统制造业原先是一种纵向格局，而现在的制造业则呈现水平格局。比如说，产品设计与生产、供应商、经销商以及用户都是在同一个平台上进行互动。
- 一个网络的虚拟社区里面是一个小世界。它有领袖、骨干成员。他们的意见常常起着主导性作用。
- 我们怎么样来正确地引导和利用信息网络为我所用、为企业所用、为社会所用，这很关键。

我们面前的两个重要课题

- 互联网已经成为现代社会基础环境的一个重要组成部分，深刻地影响着社会的政治、经济、文化和教育以及个人的工作、生活、娱乐等方方面面，在社会发展和国民经济中发挥着巨大的作用。
- 语音、文本及视频是互联网环境中信息交流和传播的最基本的三种媒介，其中言语信息（语音和对话文本）是中文信息的主要载体。
- 摆在我们面前的两个重要课题：
  - 如何利用技术手段，准确地把握舆论导向，正确地引导舆论？
  - 如何利用技术手段，让数据创造出巨大的社会和经济价值。

现状与挑战：互联网环境言语碎片的深度利用



通过互联网环境言语碎片的深度利用，有可能提高从海量复杂数据中获取知识的能力，可能加速科学与工程领域的创新步伐

美国政府将此设为国家的战略目标，2012年3月公布了“大数据研发计划”

传统的知识发现主要是少数人或团队的独立行为

挑战：

如何对互联网环境的言语碎片信息进行表示？

如何从庞杂的言语碎片中挖掘其蕴含的知识？

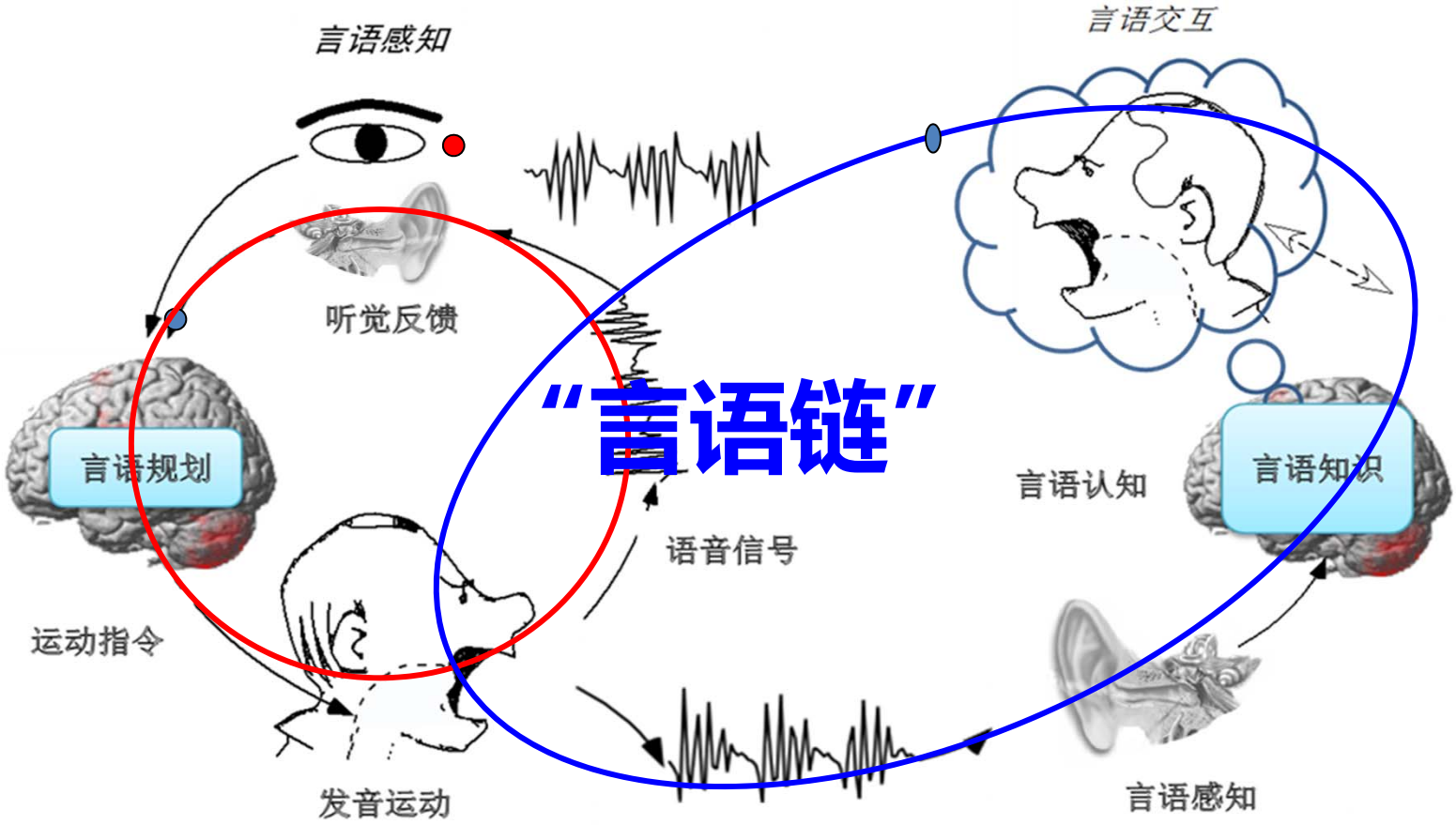
互联网规模数据处理的要求

- 理念上：大数据时代，要求我们实现信息到知识、计算到认知的转变
- 技术上：
  - 从程序到系统:与程序和算法不同，系统会对外界进行反馈。系统的行为是与平台相关的，其正确性取决于运行平台的动态特性（比如运行次数，反应时间等）。
  - 计算的理论亟需进一步扩展，从数据挖掘、信息搜索到知识获取。必须把人与系统的相互作用考虑到计算模型中来，进而不断丰富和完善计算模型。

深度计算：信息变为知识

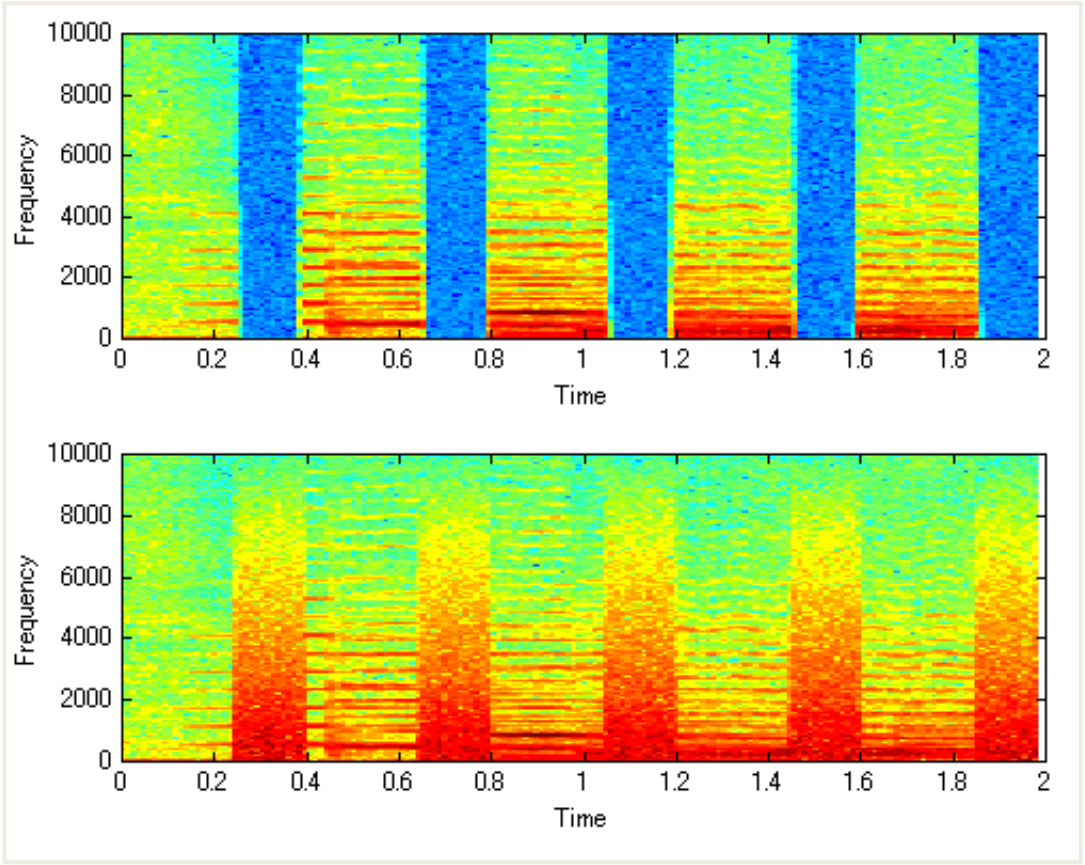
- 深度计算的基础是网络和分布服务器，所有的数据包括文本、图形、音频和视频都将成为数据。
- 深度计算的信息处理流程是数据--知识--决策，即将众多无规律的数据，通过复杂的数学模型，将数据变为知识。
- 为社会舆情分析，个人特征抽取，文化传承服务。

人的言语信息交互机制

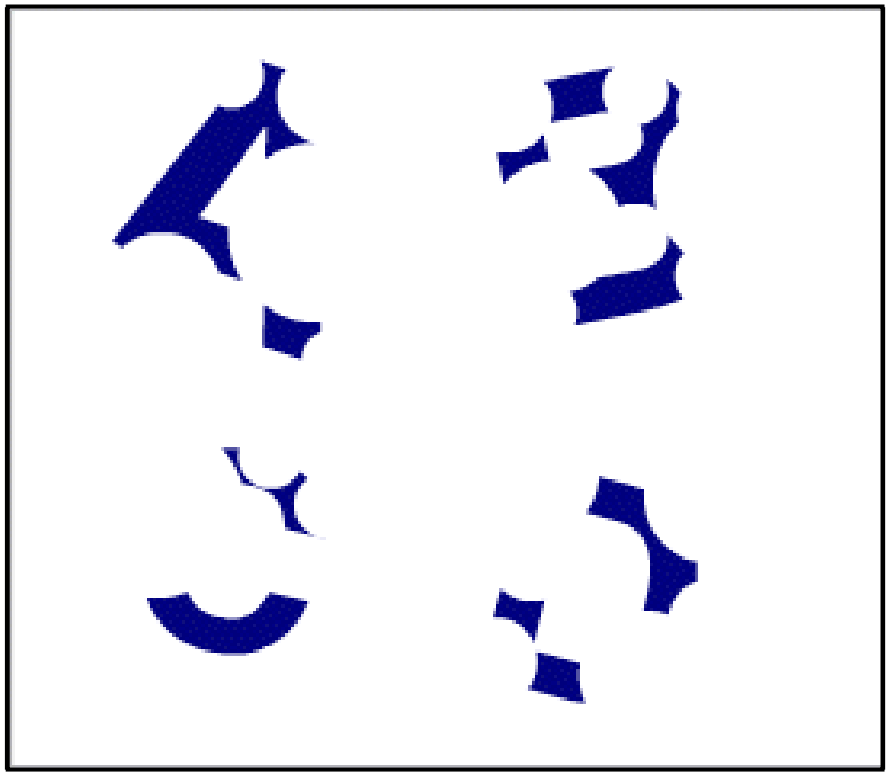


研究言语链的机理，厘清多通道感知机制，构建新的言语信息处理方法

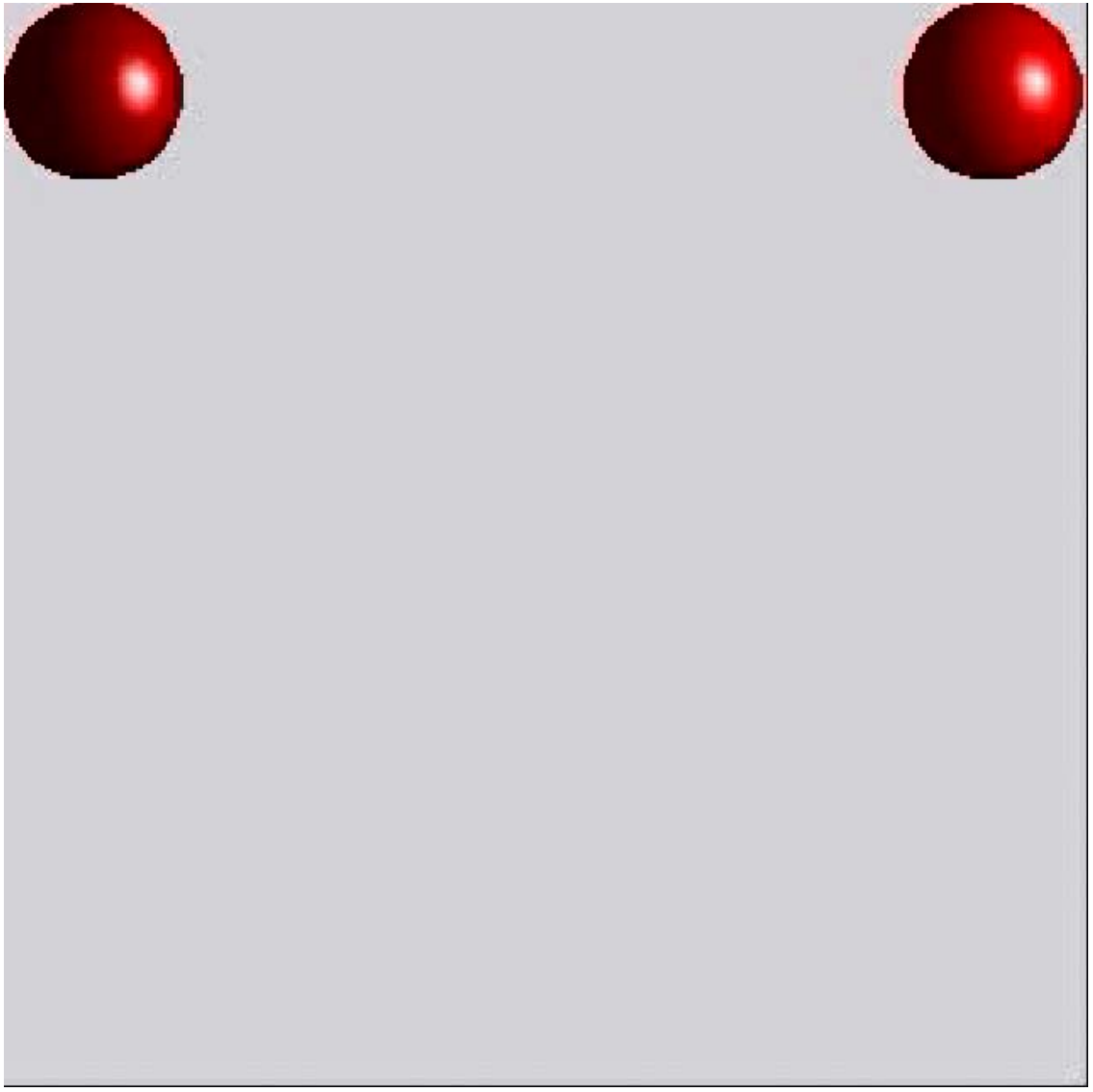
What did you listen?



What did you see?



两球的运动



What does this person say?

McGurk Effect

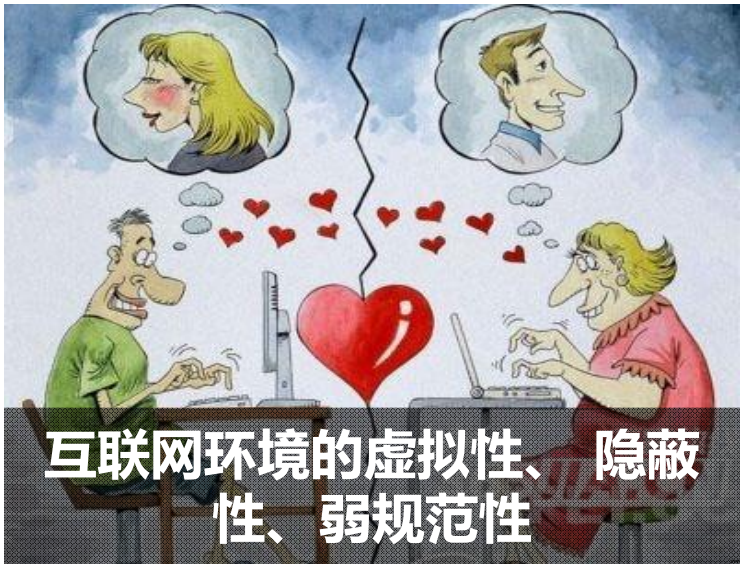
互联网环境言语信息的表示

- 言语信息表示的核心：意图表示和意图理解
- 言语意图的描述理论之一：言语行为理论

合作原则
信息适量
信息真实
话题相关
[illegible]

礼貌原则
权势要素（上下关系）
距离要素（亲疏关系）
强加要素（主从关系）

打破了传统的言语意图表示条件



挑战：
建立新的言语表示理论来描述互联网环境的言语行为

言语识别是互联网言语信息获取的关键

中文言语识别的现状：

朗读语音的识别率≈90%

即兴语音的识别率≈70%

互联网环境言语识别的困难：

- 言语口语化、信息缺省
- 多语言环境、语境多变
- 海量数据，信息碎片化
- …… 等等

如何应用人的言语处理机理，构建新的言语识别方法

互联网环境中文言语信息表示问题

互联网复杂场景中文言语多通道感知机理

互联网环境中文言语的深度理解

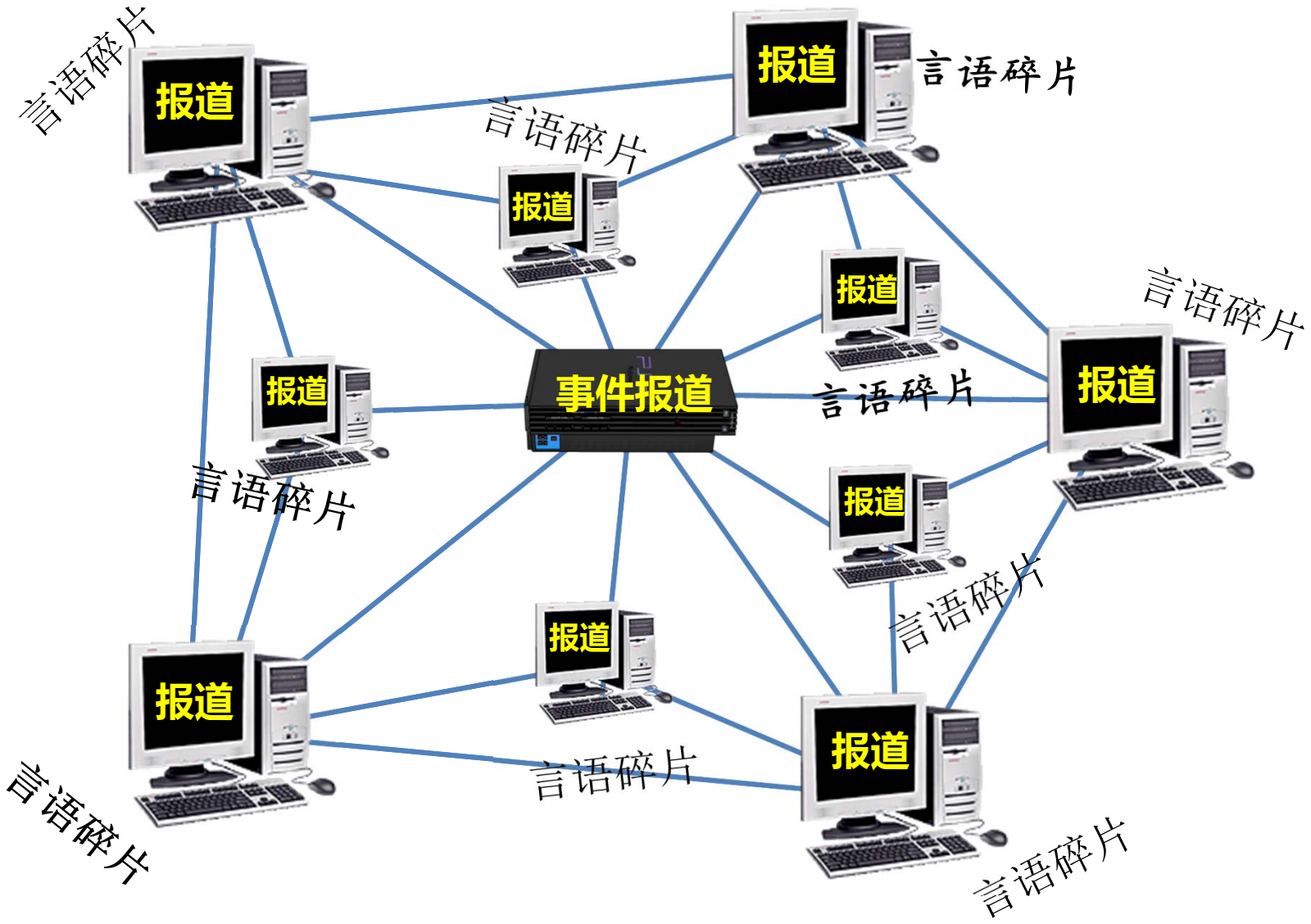
言语碎片整合与意图理解

在跨国多语言环境中，以互联网为基础，以语音通话为主要手段

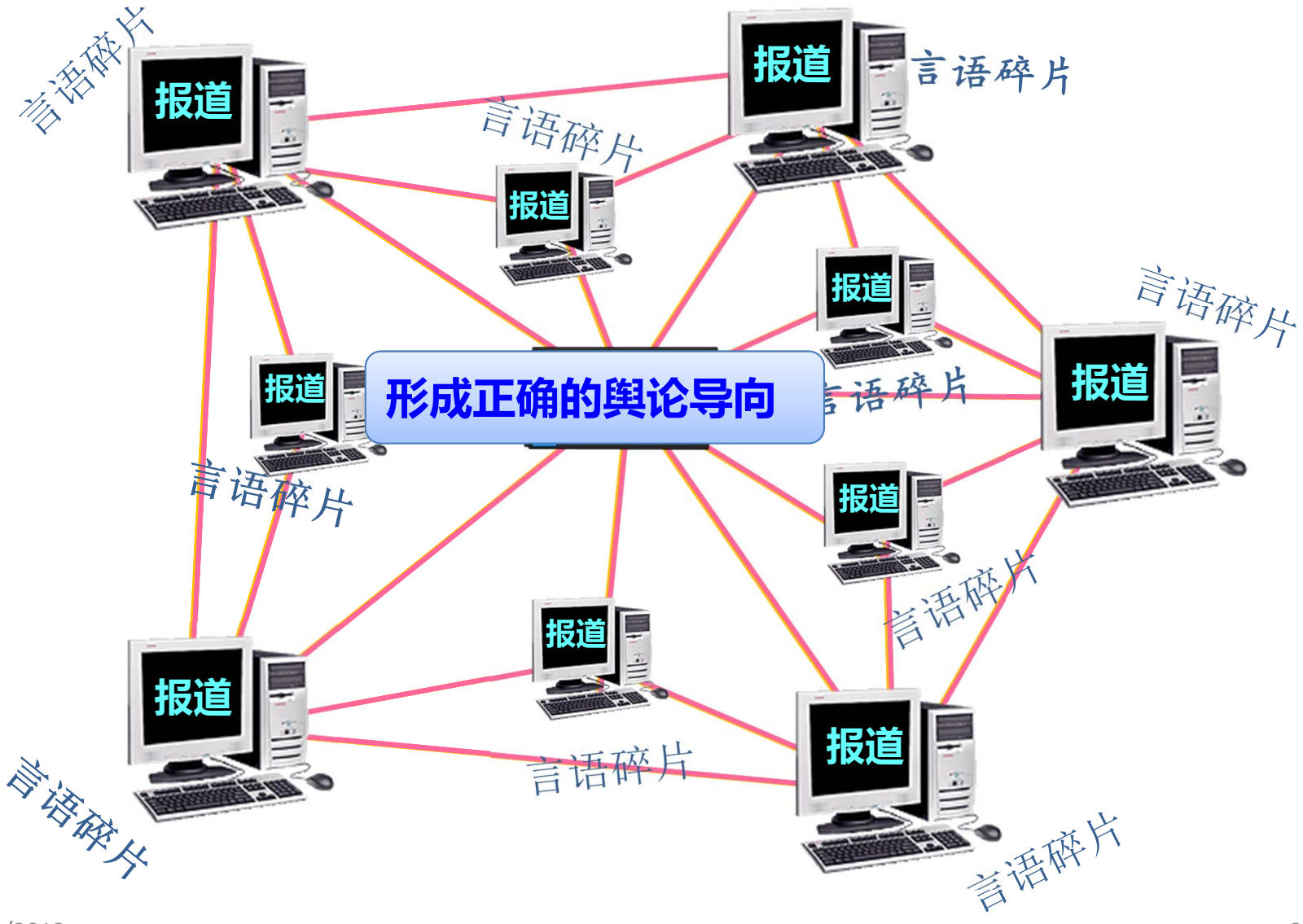
多言语识别　　机器翻译

碎片整合 → 意图理解 → 事件和目标追踪

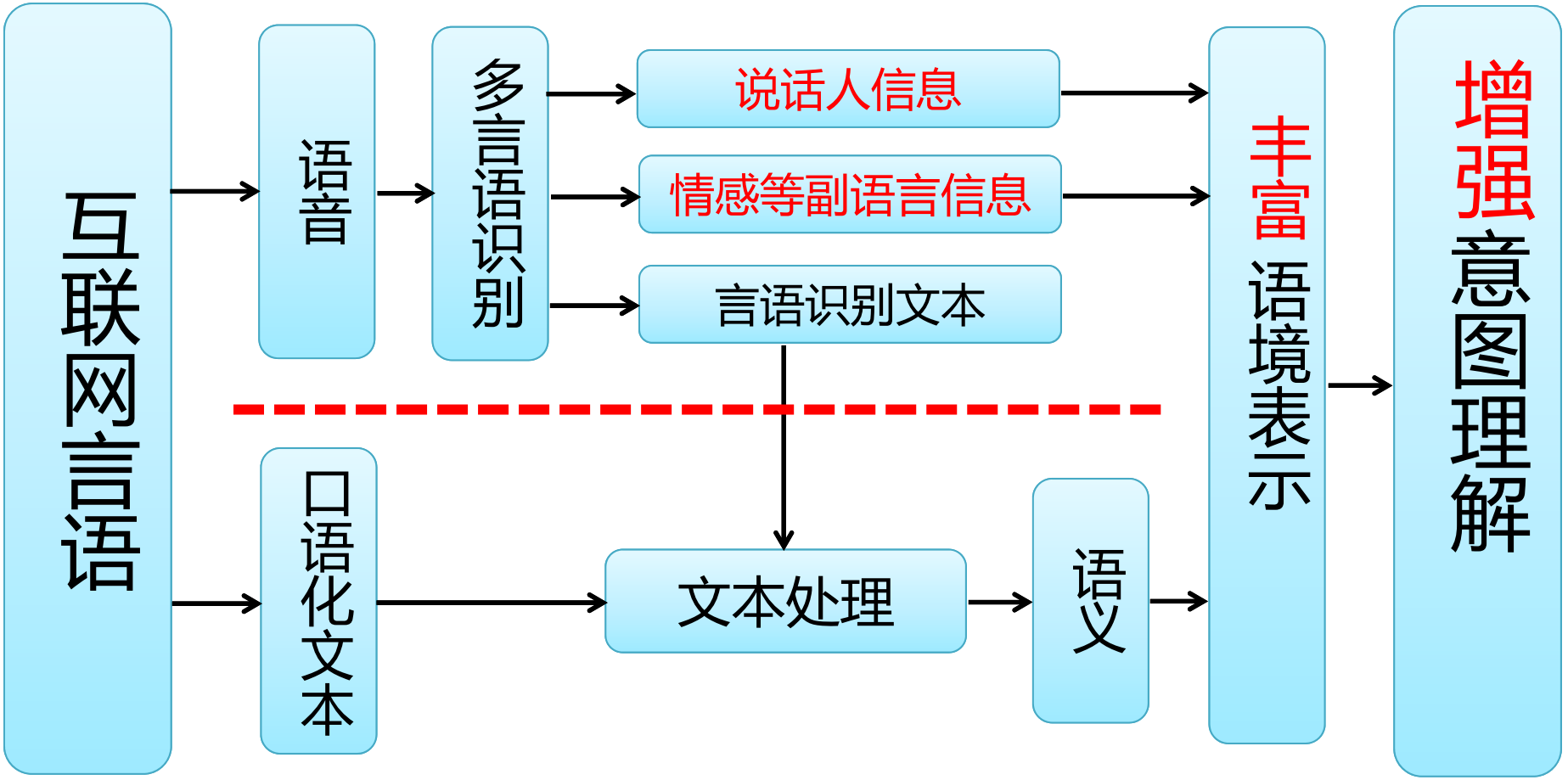
基于事件报道反馈行为分析的导向性报道



基于事件报道反馈行为分析的导向性报道

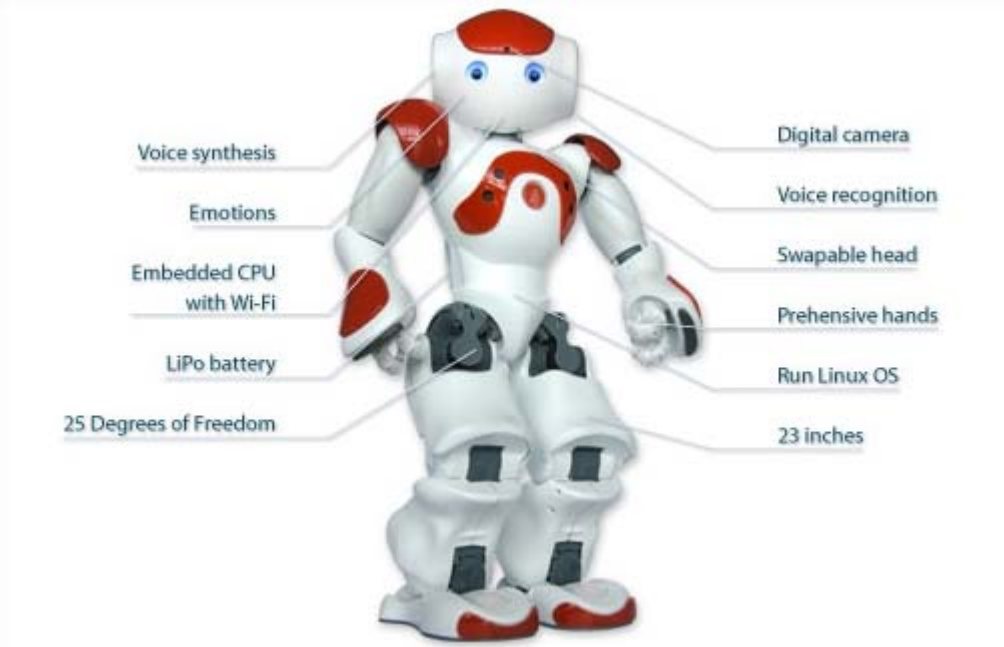


小结：多言语识别+口语化文本处理+意图理解

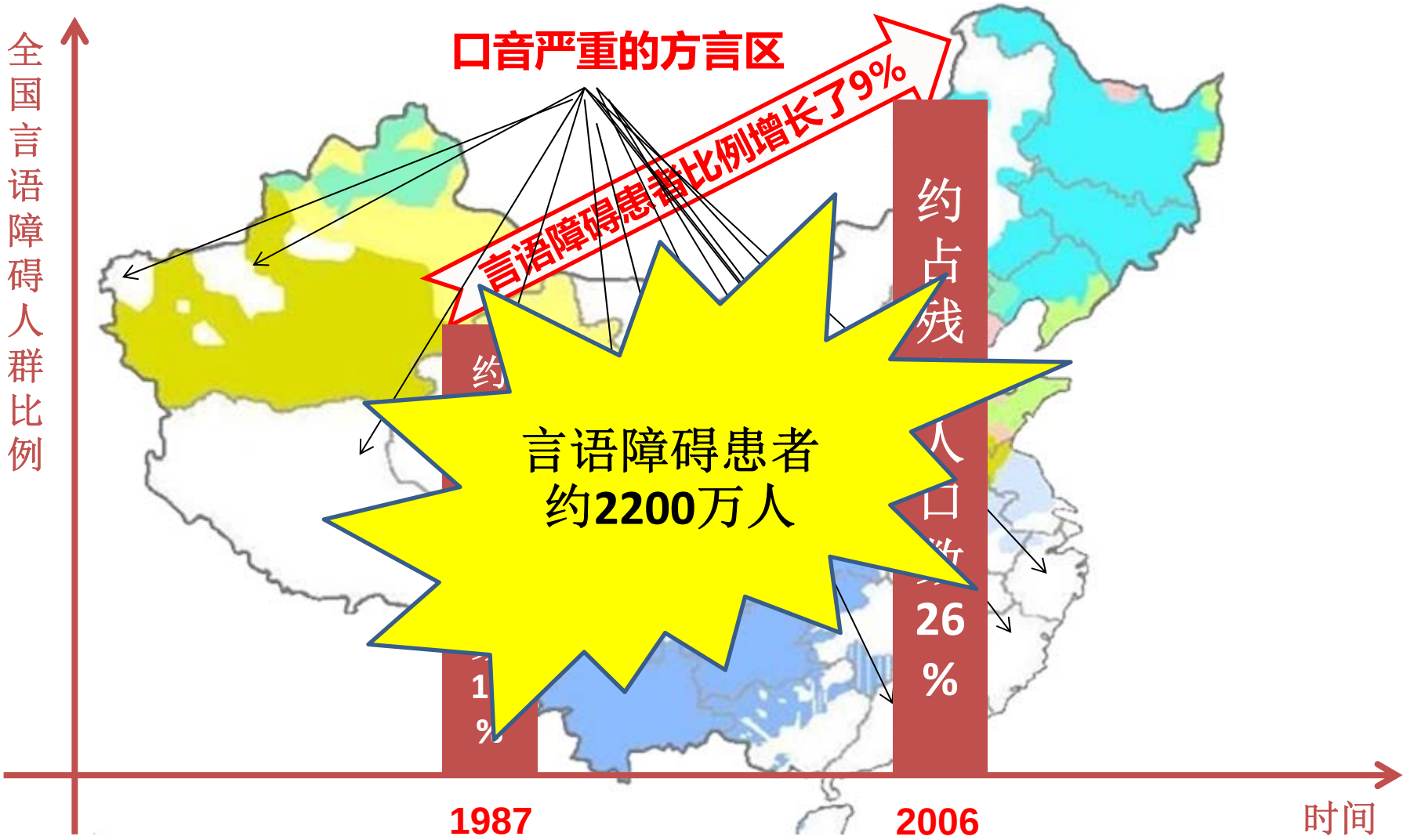


未来的社会--人机和谐社会

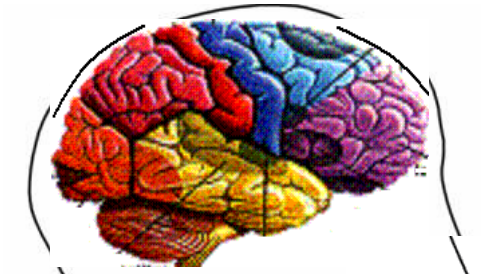
- 智能化
  - 计算机处理过程化的计算工作及事务处理工作已经达到了相当高的水平，是人力望尘莫及的。
  - 但在智能性工作方面，计算机还远远不如人脑。
  - 如何让计算机具有人脑的智能，模拟人的推理、联想、思维等功能，是一个重要发展方向。



人机和谐社会：言语障碍及言语交流障碍



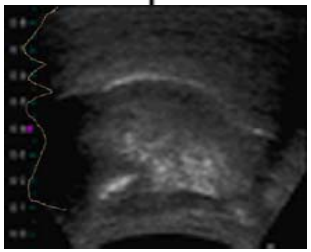
现状与问题：神经控制机理与言语康复训练

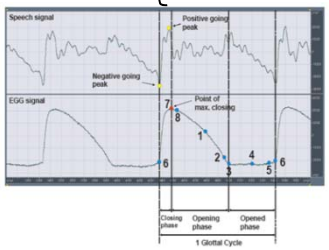


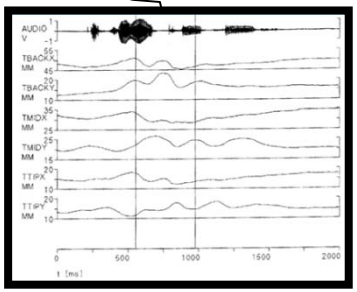
言语障碍康复训练主要依赖于医生经验

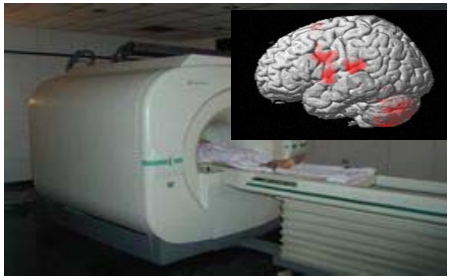
缺乏言语障碍综合分析平台和可视化的有效手段

如何构建语音产生过程的神经生理控制实验平台，通过机理探索和可视化反馈提高言语康复训练的效率

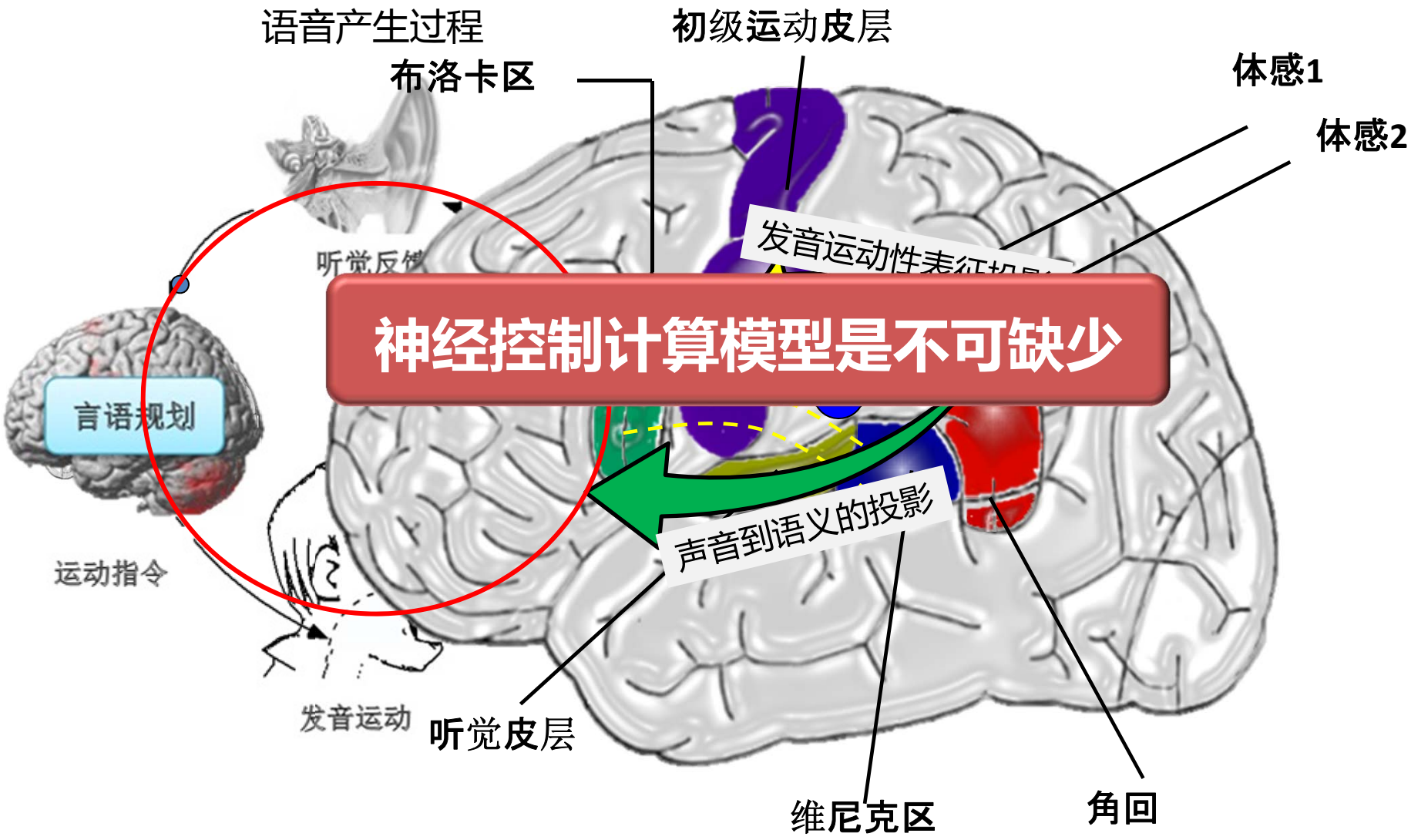




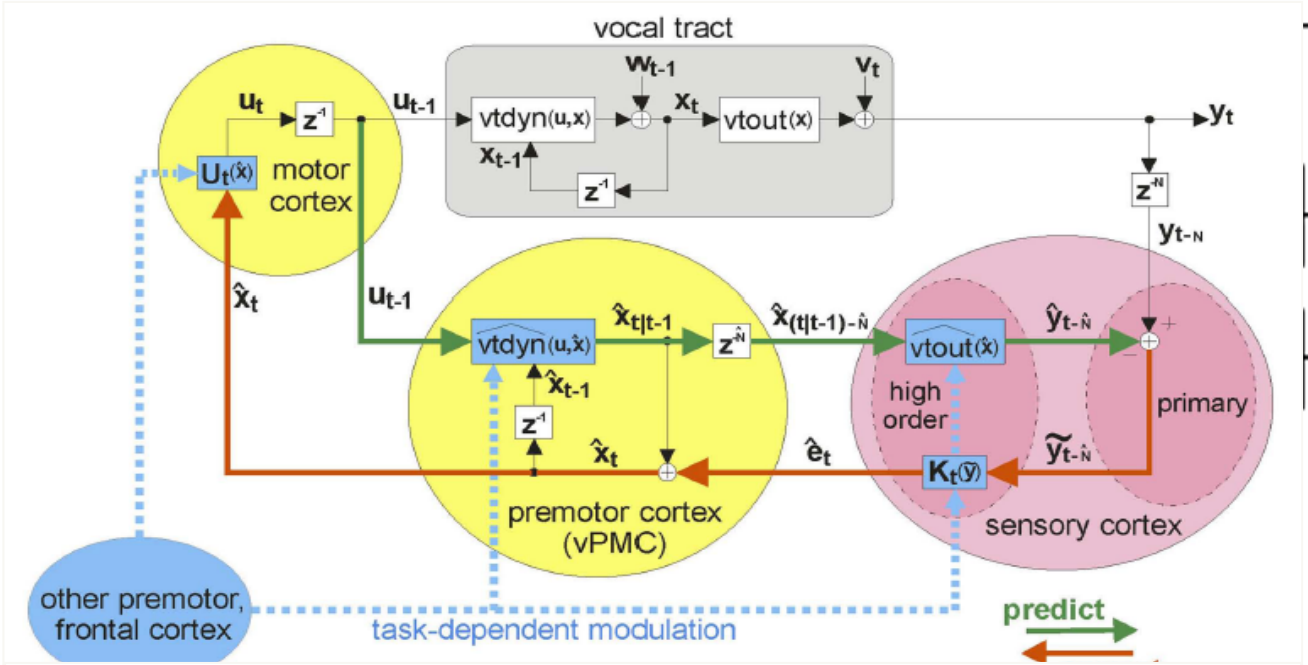


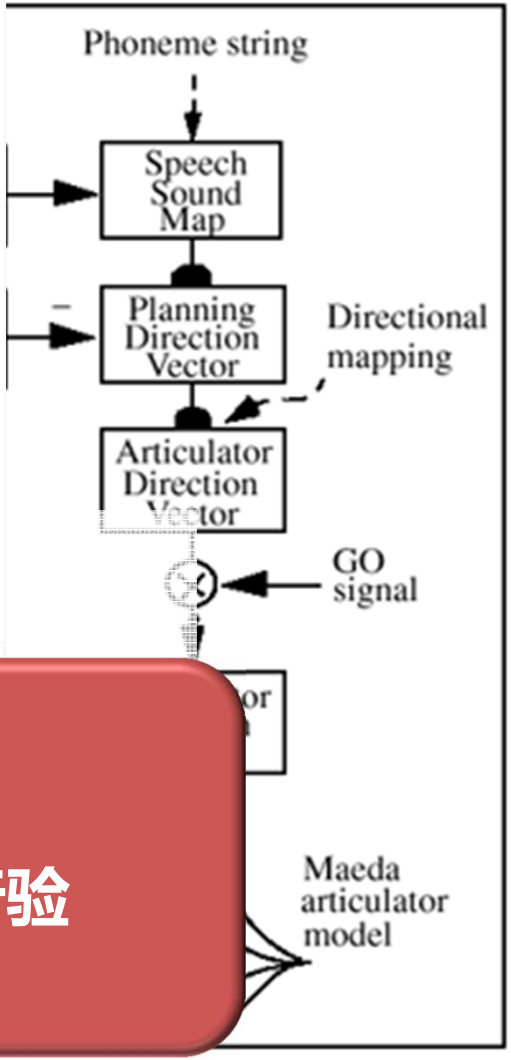


现状与问题：语音产生过程的大脑功能研究



现状与问题：语音产生神经控制机理



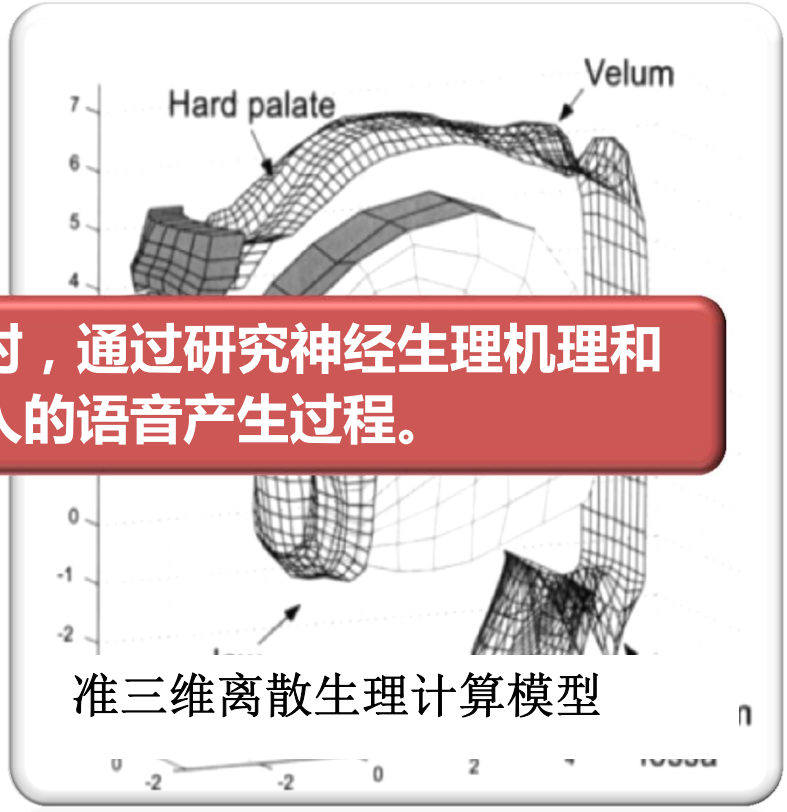


存在问题：

- **没有充分考虑神经生理学最新的发现**
- **很难从声学、生理、神经等各层次来进行验**
- **在底层缺少发音生理计算模型的支持**

现状与问题：发音生理计算模型研究

三维连续生理计算模型



准三维离散生理计算模型

在完善发音生理计算模型的同时，通过研究神经生理机理和协同控制机制，模拟人的语音产生过程。

目标及展望

言语正在成为未来互联网交际的主要媒介，以多言语识别、机器翻译和意图计算为主的深度处理和利用具有重大的理论价值和应用前景；

基于言语行为计算理论及语境模型的言语意图理解，有效地实现对互联网言语信息的深度计算，从而形成较为系统的互联网言语信息处理体系，使中文言语信息处理的理论和方法处于国际领先地位；

言语研究不仅对维护社会公共安全有重大贡献，而且对以信息内容深度理解和有效利用为主要特征的下一代ICT技术和产业产生重要影响；