智能语音技术及其应用
第一章 绪论

洪青阳

厦门大学语音实验室
http://speech.xmu.edu.cn

课程简介

语音信号处理是一门综合性学科，涉及的领域非常广泛：声学、语音学、语言学、信号处理、概率统计、信息论、模式识别、机器学习、人工智能等。

课程简介

第一部分：语音识别基础

这部分介绍基础理论，以让学生更好的了解广泛应用于现代口语对话系统的技术。这些理论包括语言学、语音学、概率理论、信息论和模式识别。这些章节是本课程后续部分的重要基础。

第二部分：语音识别关键技术

这部分深入介绍现代语音识别系统，特别是那些在建立实用系统行之有效的技术。我们还会从理论和实际角度，阐述这些技术是如何和怎样应用的。

第三部分：语音合成

这部分我们会探讨建立文本-语音系统的有关技术。合成系统实际上包含了语音识别系统的主要部分，只是他们的组合顺序正好相反。

课程主要内容

第一章 绪论，介绍人类语音的产生和感知过程、语音技术的发展历史等。
第二章 语音信号基础，介绍声音的采集和量化过程，编码和存储格式。
第三章 语音特征提取，介绍语音信号的频域分析、倒谱分析、MFCC提取过程等。
第四章 动态时间规整，介绍不等长语音的模板匹配算法。
第五章 概率统计和信息理论，介绍语音模型的必要的数学基础。
第六章 隐马尔可夫模型(HMM)，介绍双重随机过程，HMM的三大问题。
第七章 深度神经网络(DNN)，介绍深度学习在语音识别的应用。
第八章 语言模型，介绍语言模型的训练过程及其在语音识别的作用。
第九章 解码器，介绍加权有限状态机(WFST)、HCLG等关键技术。
第十章 语音合成，介绍用计算机生成语音的原理和系统构成。

课程目的及要求

- 通过本课程的学习，学生将掌握语音处理技术相关的概念、原理、方法与应用，以及该研究领域的最新进展，为日后从事工程技术工作，科学研究以及开拓新技术领域，打下坚实的基础。

- 重点内容：
  - 语音特征提取；
  - 隐马尔科夫模型(HMM)；
  - 深度神经网络(DNN)；
  - 语音识别及应用；
  - 语音合成及应用。

- 考核方法：考勤(10%)+平时作业(70%)+课程实践(20%)
- 考勤(10%)+平时作业(70%)：个人完成。
- 课程实践(20%)：可组队，建议两人一组。

参考书

- 俞栋，邓力著，解析深度学习：语音识别实践，电子工业出版社，2016年6月.
- 韩纪庆，张磊，郑铁然. 语音信号处理（第2版），清华大学出版社，2013年4月.
- Kaldi: http://kaldi-asr.org/doc/
- HTKbook: http://htk.eng.cam.ac.uk/

重要文献来源

重要期刊：

- ACM/IEEE Transaction On Audio, Speech and Language Processing
- Speech Communication
- Computer Speech and Language

重要国际会议：

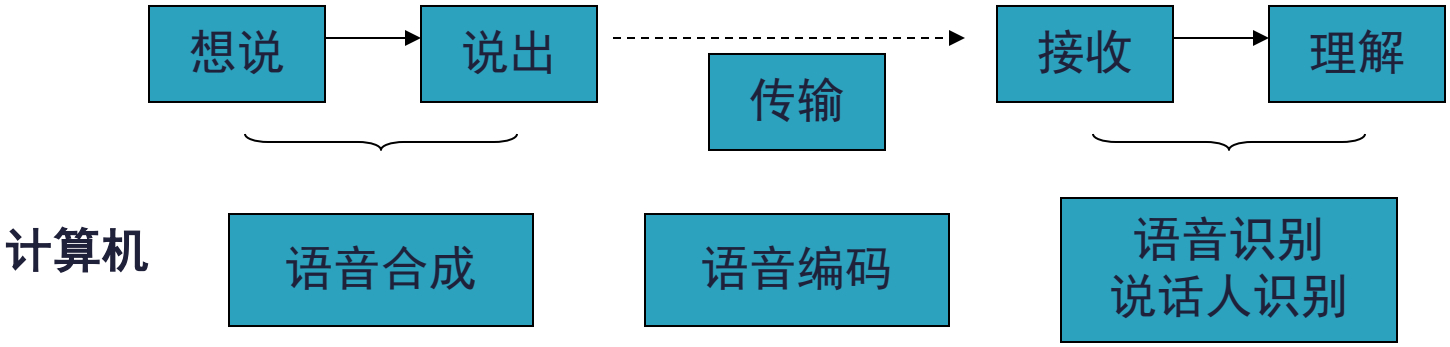
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- INTERSPEECH

语音

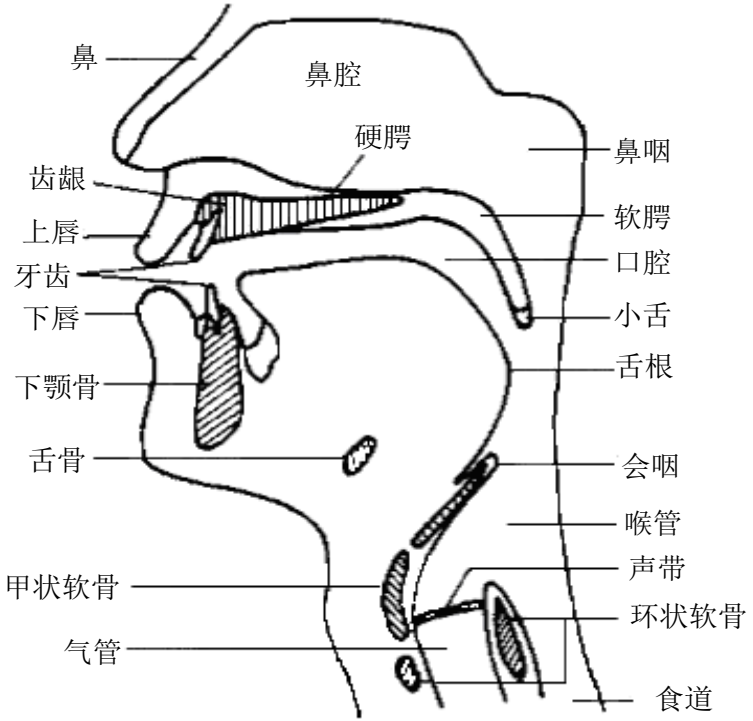
- 语音和非语音

  语音是语言的声学表现，是人类交流信息最自然、最有效、最方便的手段，是人类进行思维的依托。

- 人的言语过程



语音的产生



发音器官包括：肺、气管、喉、咽、鼻腔、口腔、唇。

声道是对发音起重要作用的器官。

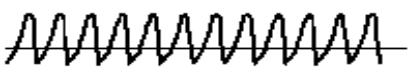
声带每开启和闭合一次的时间是基音周期（Pitch Period），其倒数为基音频率（70−450Hz）。

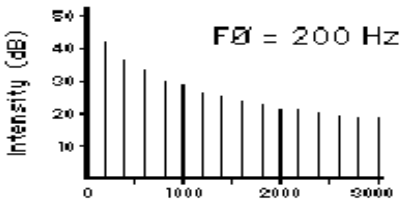
语音的产生过程

- 音源：声带音源、非声带音源
  - 声带振动周期：T（F0=1/T:基本频率）
- 声道调音：对声道形状进行调整。
  - 声道共振频率：F1、F2、F3、F4、F5
- 语音分类：
  - 浊音：由声带振动并激励声道而得到的语音。
  - 清音：由气流高速冲过某处收缩的声道所产生的语音。

语音产生模型（1）

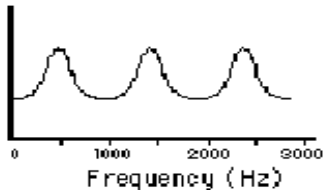
声门脉冲





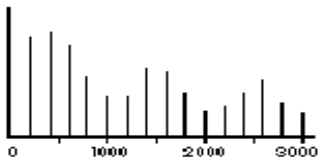
激励信号
频谱

声道



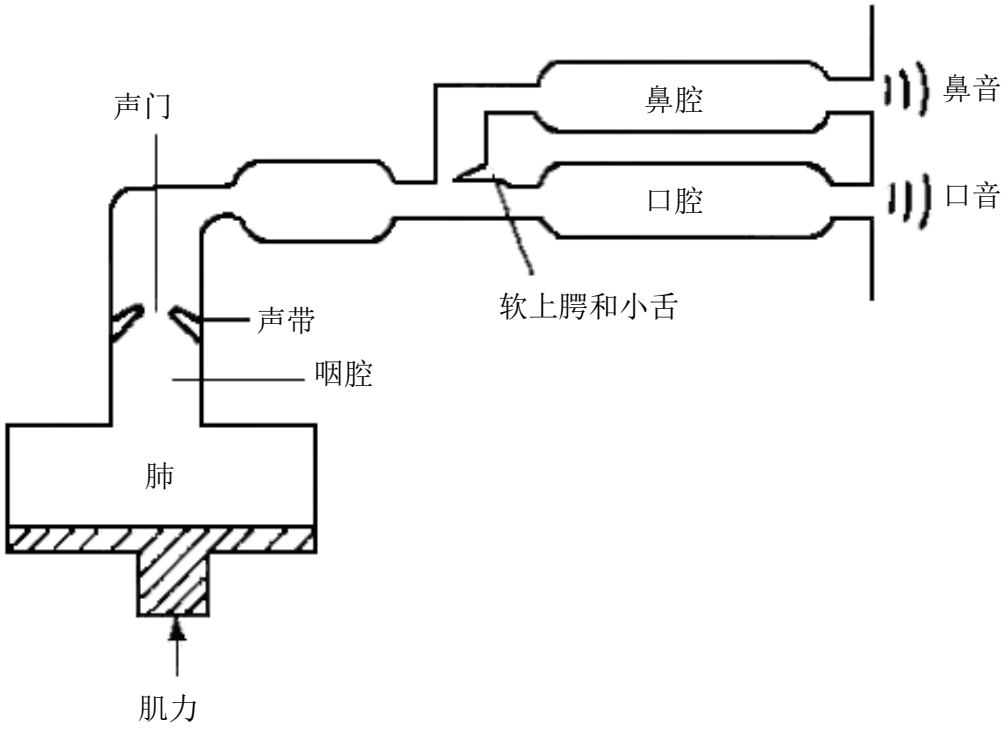
声道滤波器
传递函数

语音信号



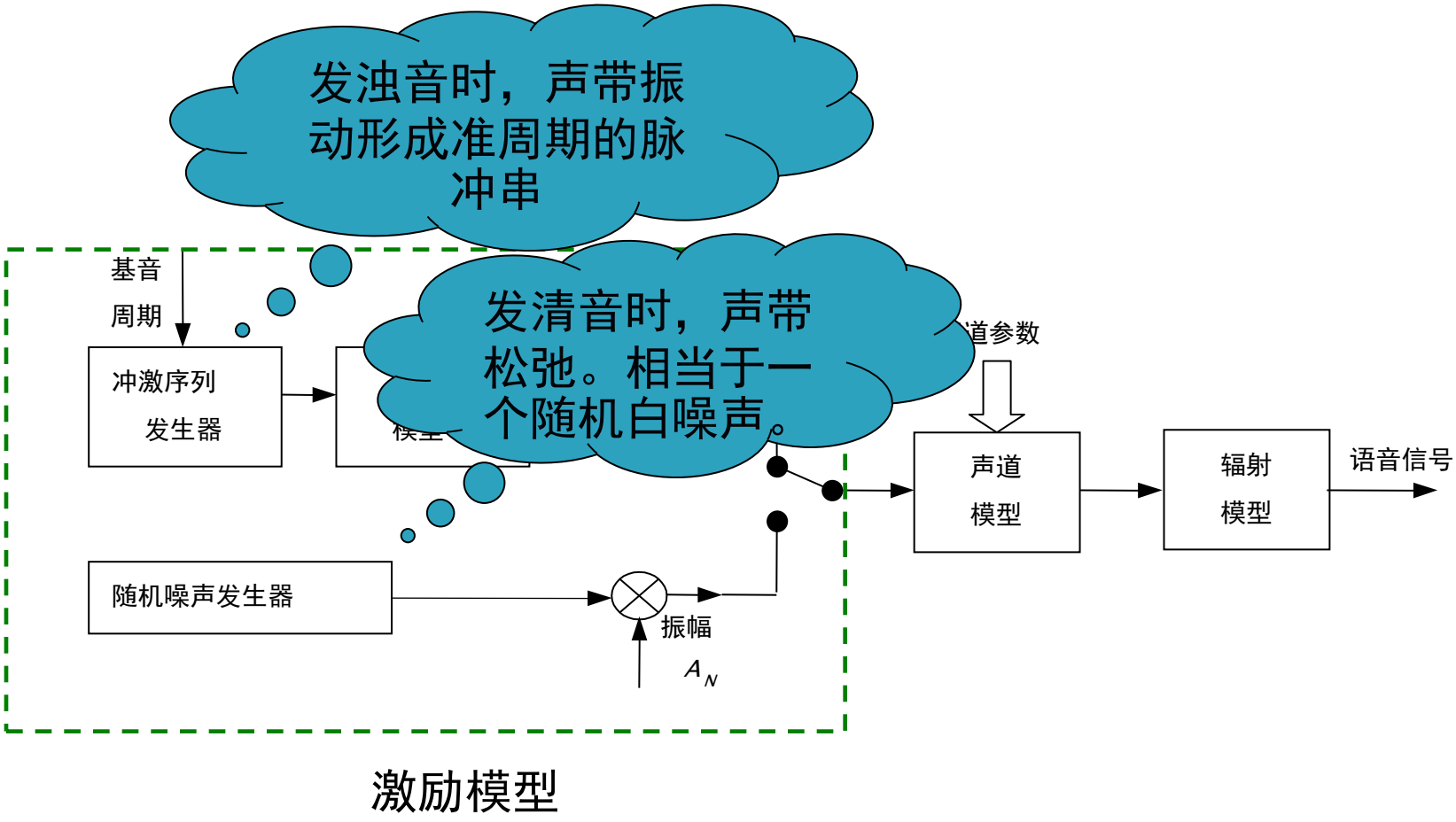
语音信号
频谱

语音产生模型（2）



语音产生的机理图

语音产生模型（3）



激励模型

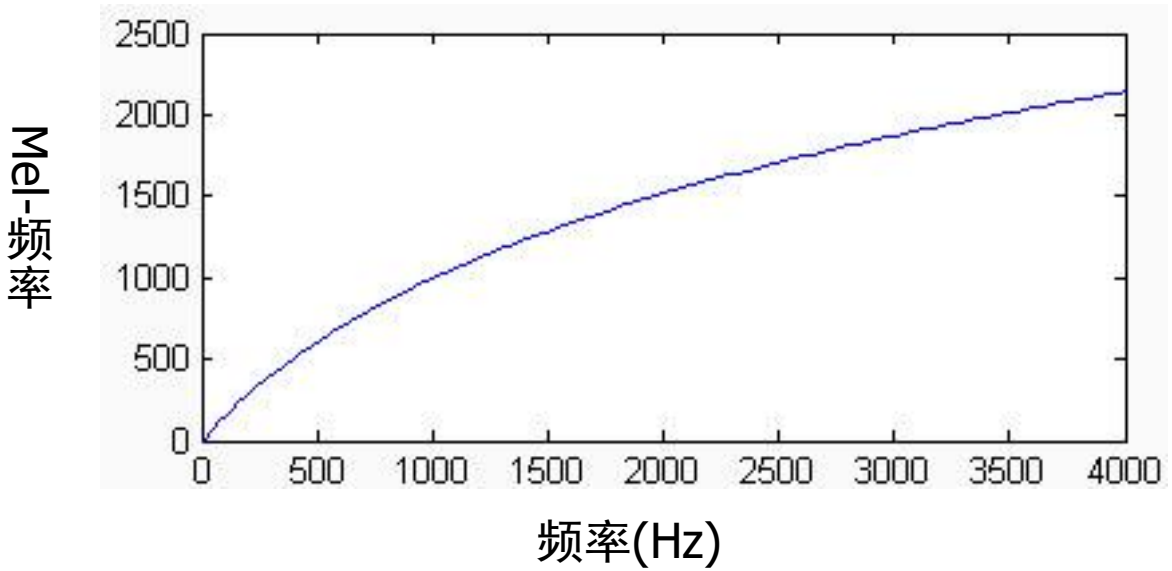
语音的感知（1）


外耳道同其
样也有共振
约是3
鼓膜位于外耳道内端的
音的
内耳。
中耳由三块
听小骨组成。
听小骨
骨
耳翼
外耳道
耳蜗神经
耳蜗
圆形窗
鼓膜
卵形窗
咽鼓管
外耳在声音感知中有两个作用，一是对声源的定
内耳的耳蜗是听觉的受纳器，从声震动到神经冲动的转换在此完成。

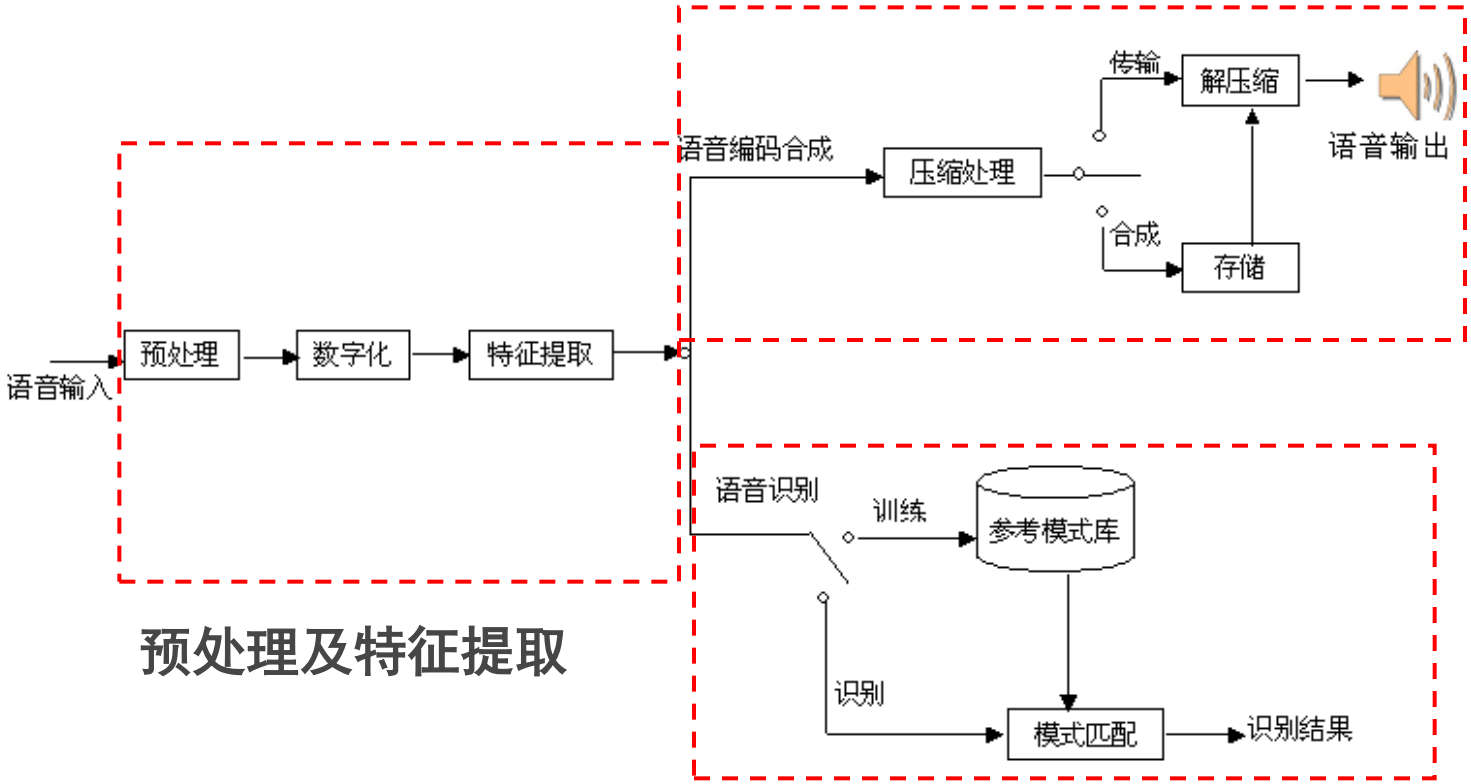
语音的感知（2）

- 正常人耳能感知的频率范围为20Hz~20KHz;强度范围为0dB~120dB。
- 音调是人耳对不同频率声音的一种主观感觉。单位为Mel，与频率近似的满足方程：

$$Mel(f) = 2595 * \log10(1+f/700)$$
$$= 1127 * \ln(1+f/700)$$



语音信号处理的总体结构框图



预处理及特征提取

智能语音技术

- 智能语音技术的分支
  - 语音识别(Speech Recognition)
  - 说话人识别(Speaker Recognition)
  - 语音合成(Speech Synthesis)
  - 语音编码(Speech Coding)
  - 语音增强(Speech Enhancement)

- 应用的需求
  - 更智能、更方便的人机接口。
  - 高质量、高效率的存储和传输。

语音技术的应用

1. 语音压缩和编码—语音通信数字化
2. 语音识别—声控应用、语音对话
3. 语音合成—自动报站、自动报时、自动警告、电话自动查询和语音提示等
4. 说话人识别—安全加密、银行信息电话查询服务以及破案和法庭取证
5. 语音增强—通常作为语音处理的前端

语音技术的发展历史

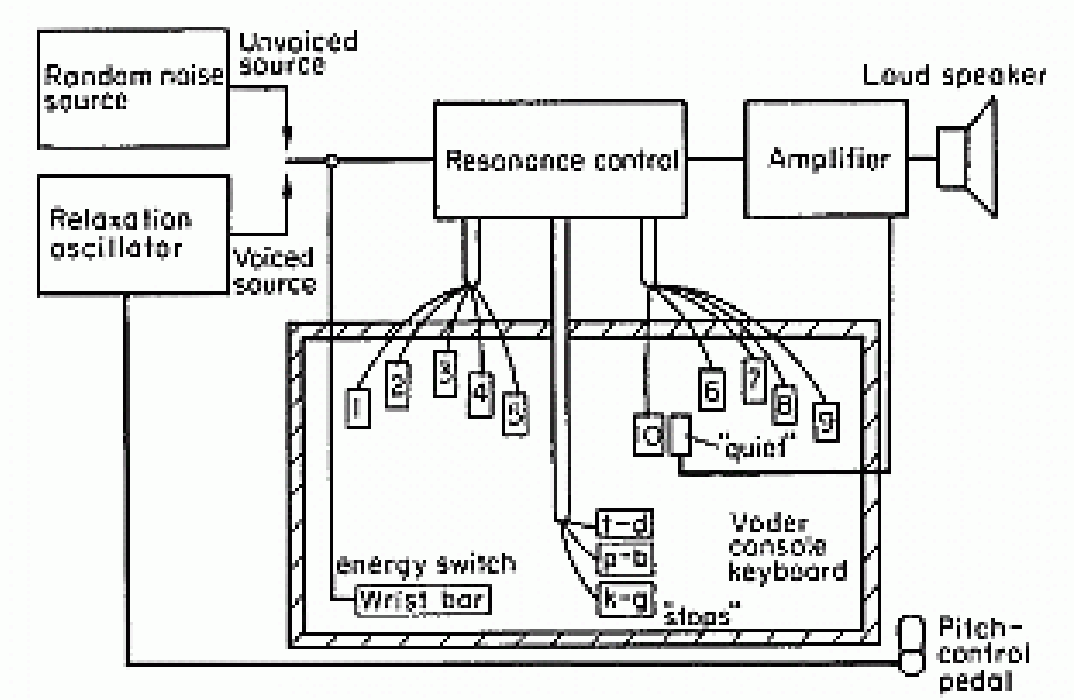
1876年Bell发明电话。

Alexander Graham Bell
(1847-1922)

Bell（贝尔）的早期电话装置

语音技术的发展历史

1939年H.Dudley研制成功第一个声码器。



A block schematic of Homer Dudley's VODER

B.H. Juang# & Lawrence R. Rabiner. Automatic Speech Recognition – A Brief History of the Technology Development, 2004

语音技术的发展历史

1942年Bell实验室发明了语谱仪。
1948年美国Haskin实验室研制成功“语图回放机”。
1952年Bell实验室研制成识别十个英语数字识别器。
1956年声控打字机。

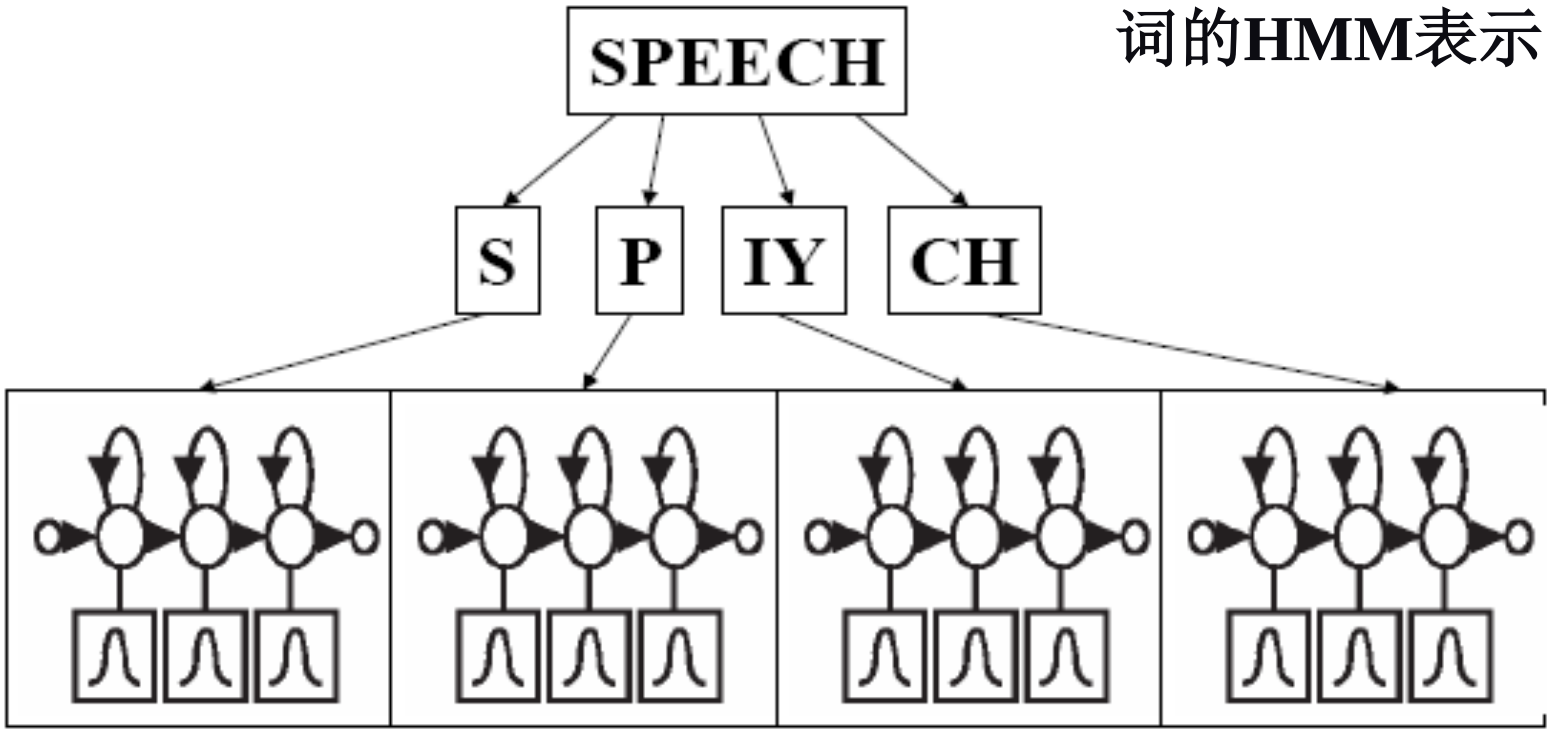
60年代以后，随着计算机技术的发展，语音信号处理技术获得了长足的进步，计算机模拟实验取代了硬件研制的传统做法。各种突破性的思想不断涌现。

语音技术的发展历史

语音识别方面：

- 1960年Denes等人用计算机实现自动语音识别，引入了时间归正算法改进匹配性能；
- 60年代中期，Martin等人为邮局研制了邮政编码阅读机；
- 70年代起，人工智能技术开始引入到语音识别中。美国国防部ARPA组织了有CMU等五个单位参加的一项大规模语音识别和理解研究计划；
- 70年代中，日本学者Sakoe提出的**动态时间规整(DTW)**算法对小词表的研究获得了成功，从而掀起了语音识别的研究热潮；

语音技术的发展历史—语音识别



词的HMM表示

词汇达到了20000，识别率达到了94. 6%。

语音技术的发展历史—语音识别

- 90年代初，CMU的Lee Kaifu完成的非特定人连续语音识别系统SPHINX是最有代表性的，它能识别997个词汇的连续语音，识别率达到95.8%；
- 1997年IBM推出的汉语听写机产品ViaVoice为语音识别在汉字输入方面的实际应用开辟了新的道路；
- 1999年Intel推出语音识别软件开发包Spark3.0；
- Microsoft发布Speech SDK语音识别引擎。

语音技术的发展历史—语音识别

- 90年代至今，语音处理技术产品化，面向个人用途的连续语音听写机技术也日趋完善。国内清华大学、中科院声学所和中科院自动化所在汉语听写机研究方面有一定成果。

- 2011年，苹果推出支持自然语音理解功能的iPhone4S-Siri。通过整合人工智能分析系统和大量网络服务的APIs，Siri跟原有的单一语音控制软件拉开了功能应用上的等级差，给全球用户带来了耳目一新的感觉。

语音技术的发展历史—语音识别

- **深度学习是近年来在线语音识别的一个重大突破，基于云端深度学习算法和大数据库支撑的在线语音识别率目前可以做到95%以上，科大讯飞、百度都提供了达到商业标准的语音识别服务。**

语音云

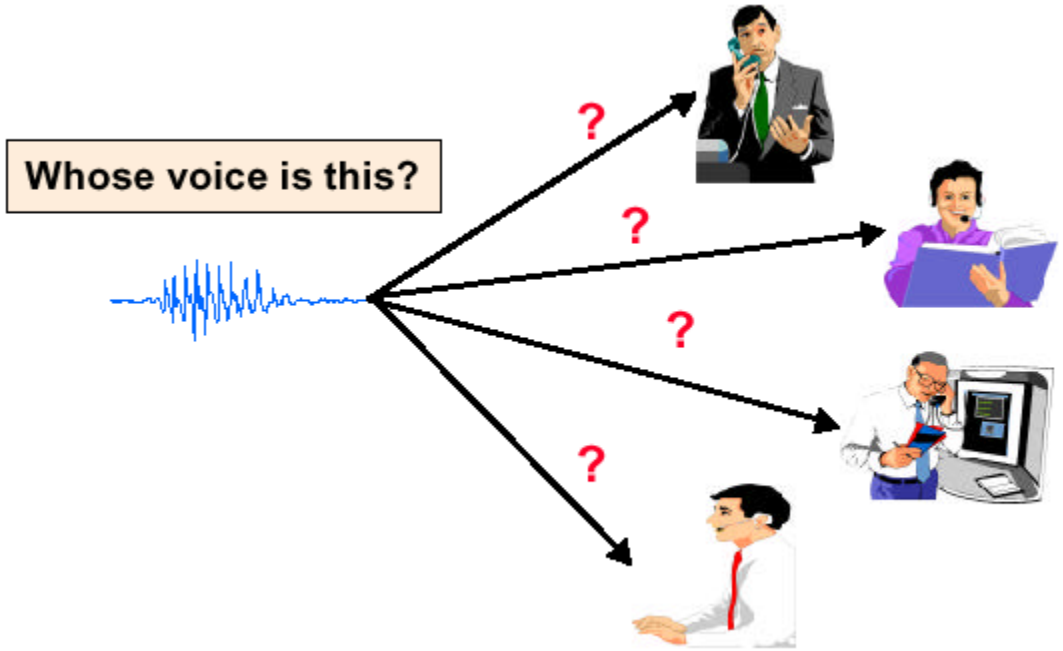
语音输入法

语音搜索

语音技术的发展历史

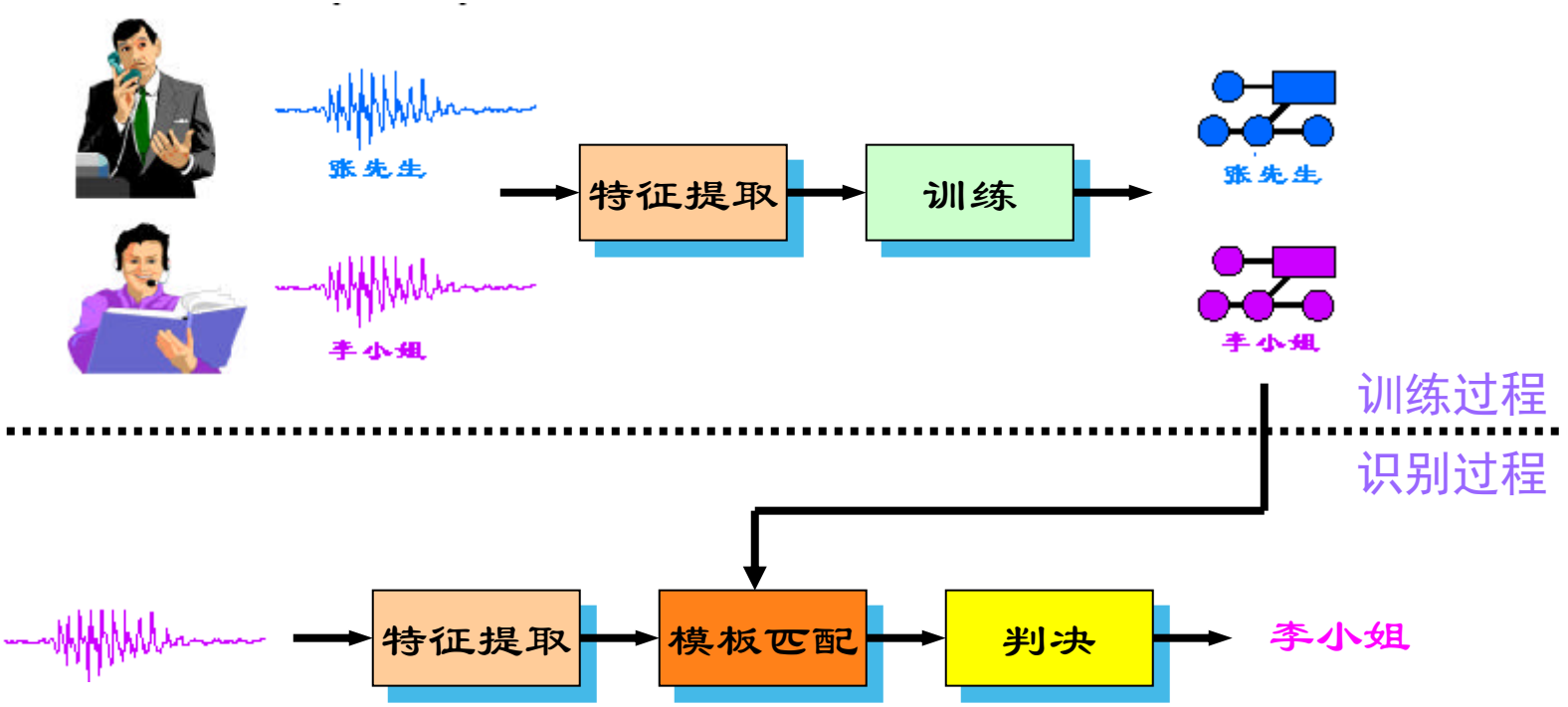
说话人识别(又称**声纹识别**)方面：

- 每个人的发音器官和发音习惯都有差异。通过对语音信号的分析和处理，提取代表说话人个性信息的特征，计算机就能够自动地鉴别说话人的身份。



语音技术的发展历史—说话人识别

- 说话人识别的基本原理



语音技术的发展历史—说话人识别

- 最早的研究是在第二次世界大战期间，美国国防部向贝尔实验室提出课题，要求根据窃听的电话语音判断说话人是哪一位德军将领。
- 二十世纪七十年代以前由于计算技术的限制，研究进展缓慢，DTW为主要算法。
- 八十年代后获得了突飞猛进的发展，出现了各种算法，文本相关的系统达到了可以商业化的水平，文本无关的系统成为研究热点。
- 在伊拉克战争期间，美国FBI宣称在电视台发表讲话的不是萨达姆本人，而德国科学家应用说话人识别技术证实讲话人确实是萨达姆。
- 九十年代以来，GMM-UBM, SVM, i-vector先后成为主流算法。

语音技术的发展历史—说话人识别

两个主要应用领域：

- 身份鉴别：使用说话人识别技术判断使用者是否有权利进入相应的场所，或访问相应的信息。
- 国防、刑侦领域的话者信息检索。

2001年迫降在我国海南机场的美军EP-3侦察机中就装有声纹侦听模块。

语音技术的发展历史

语音编码方面：

- 七十年代起，国外就开始研究计算机网络上的语音通信，主要是基于ARPANET网络平台进行研究；
- 1974年首次分组语音实验是在美国西海岸南加州大学和东海岸的林肯实验室间进行，数码率为9.6kb/s；
- 1975年1月美国实现使用LPC声码器的分组语音电话会议 ；
- 八十年代的研究集中在局域网上的语音通信，最早的实验是由英国剑桥大学于1982年在10Mb/s的剑桥环形网上进行的；

语音技术的发展历史—语音编码

- 1988年美国公布了一个4.8kb/s的码激励线性预测编码（CELP）语音编码标准算法；
- 进入九十年代，随着Internet的兴起和语音编码技术的发展，IP分组语音通信技术获得了突破性的进展。如网络游戏中，语音聊天，IP电话技术；
- 九十年代中期还出现了很多被广泛使用的语音编码国际标准，如数码率为5.3/6.4kb/s的G.723.1、数码率为8kb/s的G.729等 。

语音编码产品化的过程相对来说比语音识别容易

语音技术的发展历史

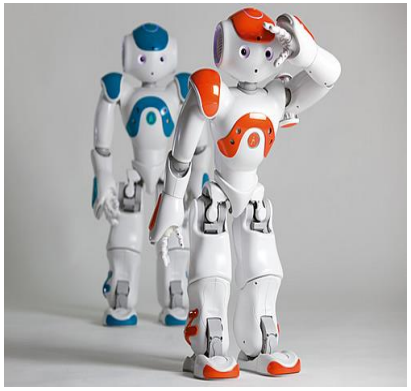
语音合成方面：

- 七十年代末开始，出现了文-语转换系统的研究，其特点是采用最基本的语音单元 ；
- 八十年代，由D.Klatt设计的串/并联混合型共振峰合成器是比较有代表性的工作；
- 九十年代末，日本的研究者提出了一种多样本、不等长语音拼接合成技术PSOLA。
- 2005年以后，以统计参数合成为主，HMM是最常用的统计声学模型形式。合成语音整体非常平稳，自然度高。
- 近年来，更多的采用深度学习方法。

人工智能和机器人

- 比尔盖茨早在2007年2月就在《科学人》杂志预言：机器人将成为下个热门领域，对工作、通信、学习与娱乐带来冲击，就像个人电脑过去30年来的影响一样。
- 人工智能和机器人产业是最有可能实现1万亿元产值的新兴行业。
- 智能机器人将走进千家万户。预测在2035年，每5人拥有1台机器人。

智能机器人



语音智能机器人(识别+理解)

智能家居

- 智能语音与智能家居的融合是大势所趋

MIT Technology Review预测，2020年全球智能设备数量将会达到280亿部，智能家庭将启动一个新的万亿级市场。未来，智能电视、智能汽车、可穿戴设备、智能空调与冰箱等将全面覆盖我们的生活。

亚马逊Echo音箱

内置Alexa语音服务(AVS)

总结

- 语音技术包括语音识别、说话人识别、语音编码、语音合成等部分。
- 经过几十年的发展，大部分语音技术已达到或接近实用水平。
- 语音技术是人机交互的关键技术，智能机器人、智能家居、智能音箱是应用热点。

Thank you!

Any questions?