中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

语音交互技术
——绪论

中国科学院自动化研究所
模式识别国家重点实验室

陶建华
jhtao@nlpr.ia.ac.cn

课程基本情况

- 普及课， 40学时
- 预修课程：机器学习基础、概率论与数理统计、数字信号处理
- 教学目的和要求：
  - 通过本课程的学习，希望学生能了解本领域研究的历史、现状、趋势和主流技术的理论和方法，掌握语音信号处理、语音识别与合成、语音对话、语音转换等概念和方法，为进一步研究或应用语音技术打下基础。

课程大纲

- 第一章 绪论
- 第二章 语音信号处理
- 第三章 算法基础
- 第四章 语音识别
- 第五章 语音合成
- 第六章 语音增强
- 第七章 语音转换
- 第八章 声纹识别
- 第九章 情感语音
- 第十章 语音对话系统

语音技术特点

语音技术是研究用数字信号处理技术和机器学习方法对语音信号进行处理的一门学科。

涵盖信号处理、自然语言处理、机器学习、信息论等多种不同的学科。

授课特点和形式

- 课堂讲述和课后练习相结合
- 讲授内容既包含传统内容，也注意吸收最新研究成果
- 既考虑一般学生普及入门的需求，也考虑相关专业学生更高的要求
- 考核方法：大作业+开卷考试

参考资料

1. 俞栋，邓力，解析深度学习：语音识别实践，电子工业出版社，2016
2. 赵力等，语音信号处理，机械工业出版社，2016
3. L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.
4. 鲍怀翘，林茂灿等，《实验语音学概要》，北京大学出版社，2014。
5. Paul Taylor, Text-to-Speech Synthesis, Cambridge University Press, 2009

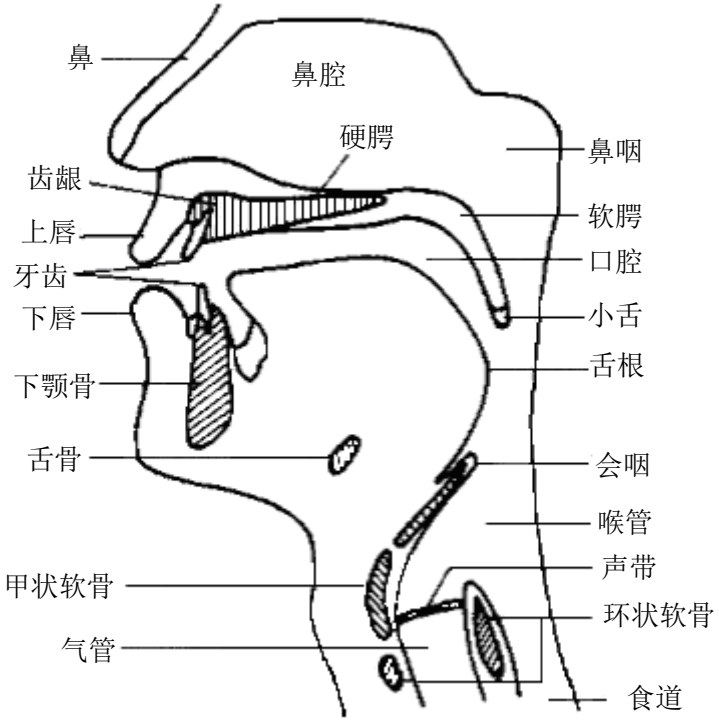
本节课提纲

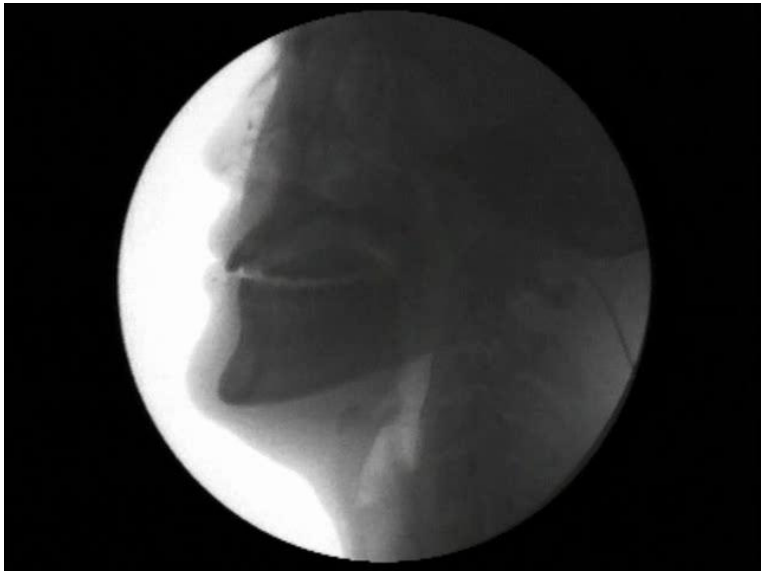
- 语音基本概念
- 语音研究历史
- 语音技术概述

本节课提纲

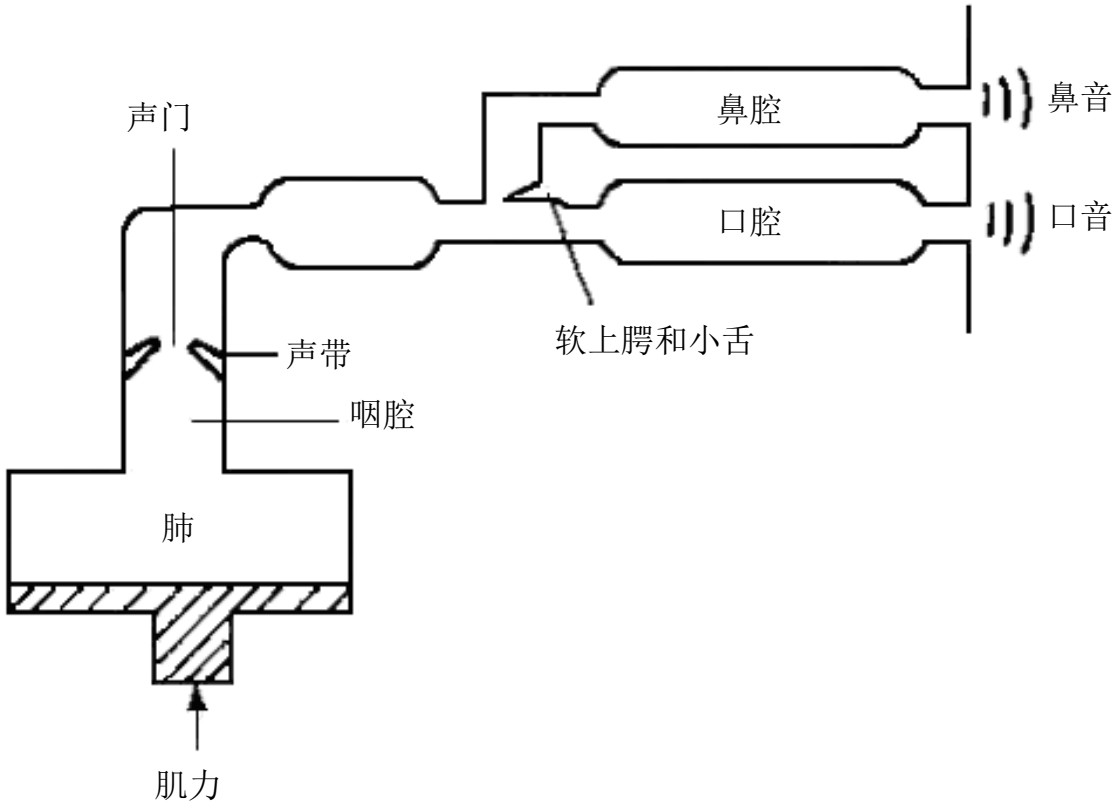
- 语音基本概念
- 语音研究历史
- 语音技术概述

语音产生的过程及声学特征

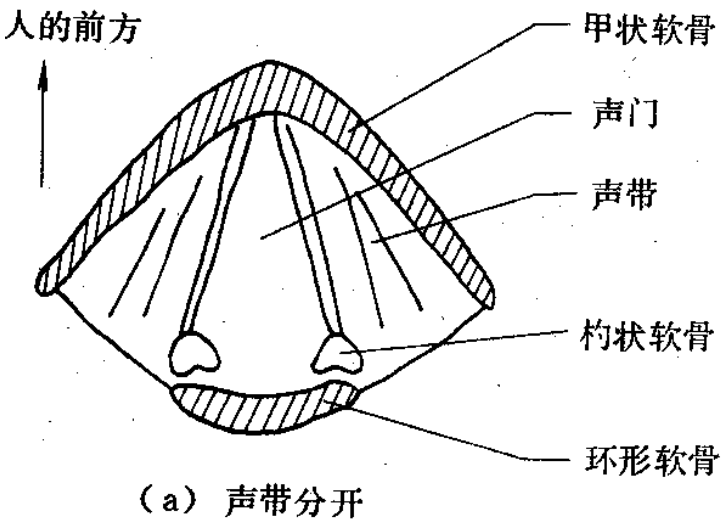




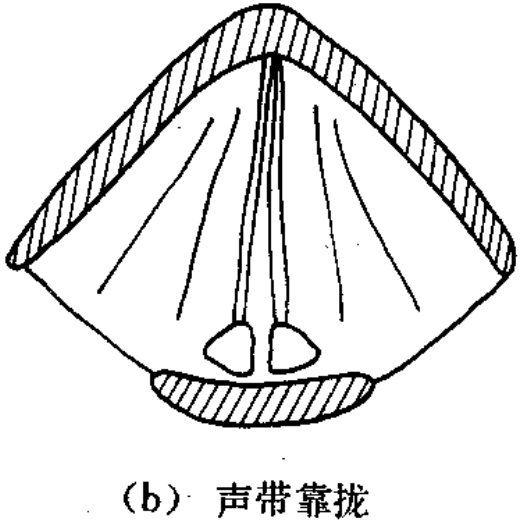
发音器官的机理模型



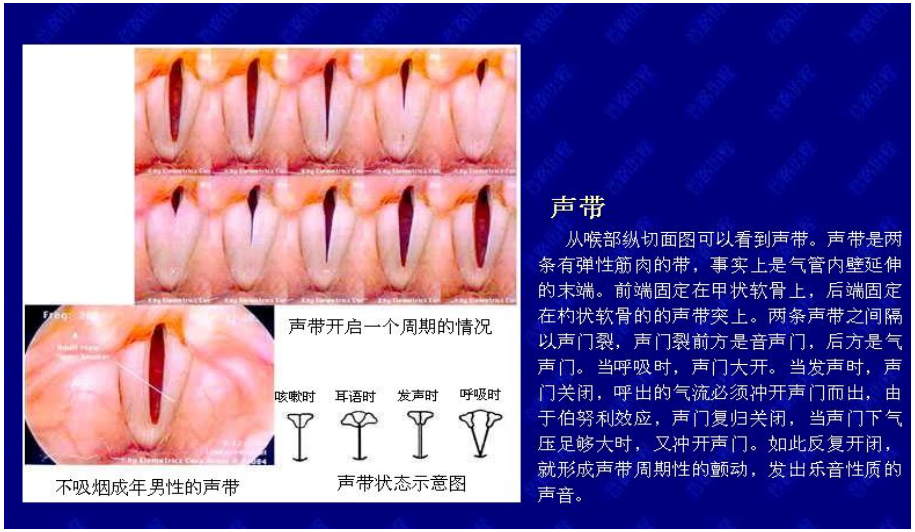
喉



（a） 声带分开



（b） 声带靠拢



声带

从喉部纵切面图可以看到声带。声带是两条有弹性筋肉的带，事实上是气管内壁延伸的末端。前端固定在甲状软骨上，后端固定在杓状软骨的的声带突上。两条声带之间隔以声门裂，声门裂前方是音声门，后方是气声门。当呼吸时，声门大开。当发声时，声门关闭，呼出的气流必须冲开声门而出，由于伯努利效应，声门复归关闭，当声门下气压足够大时，又冲开声门。如此反复开闭，就形成声带周期性的颤动，发出乐音性质的声音。

声门脉冲

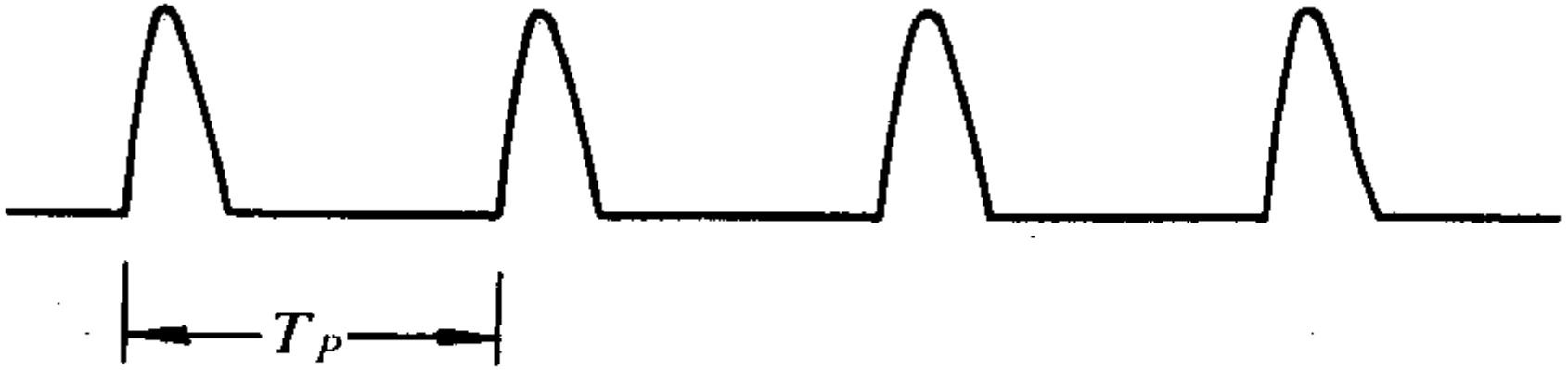
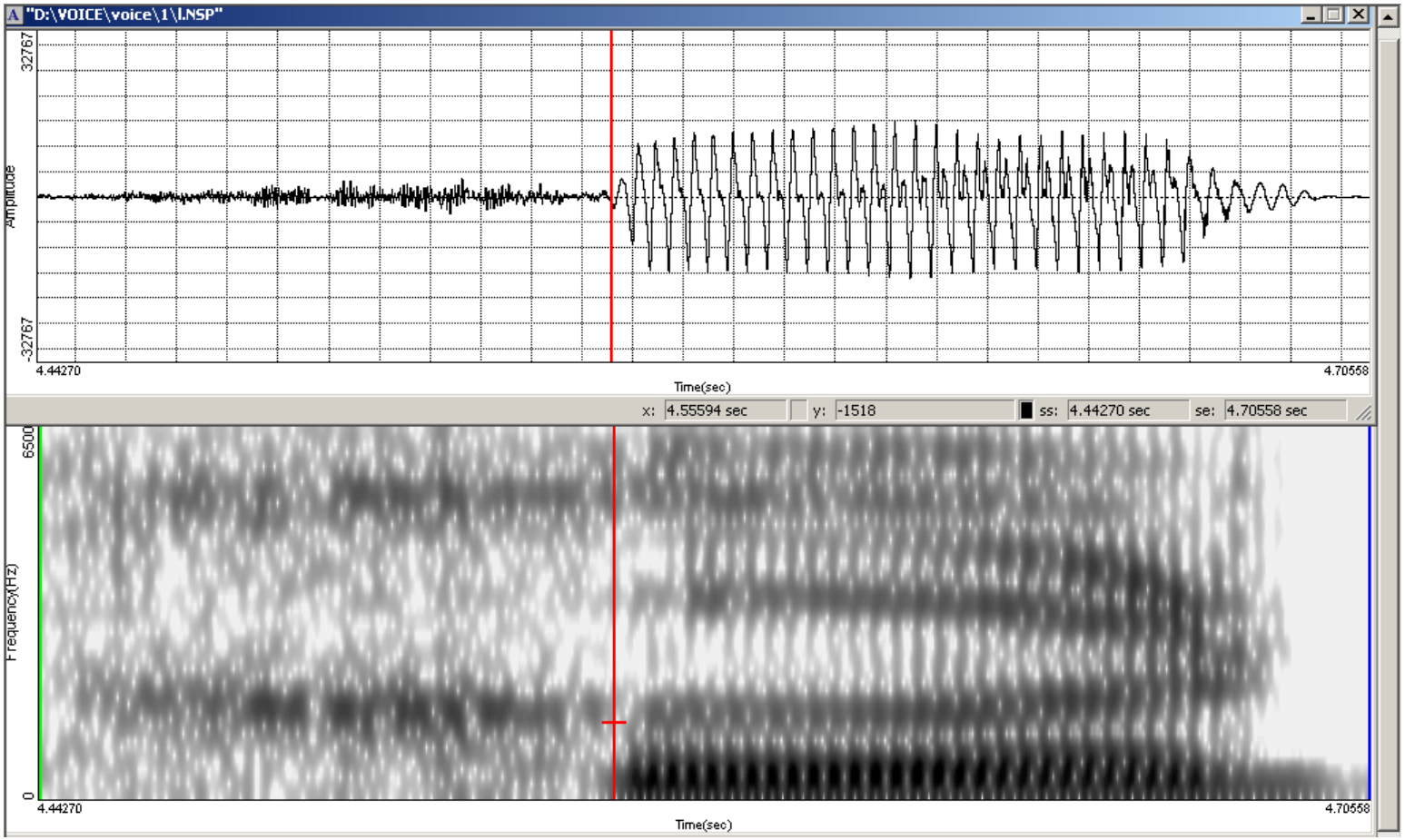
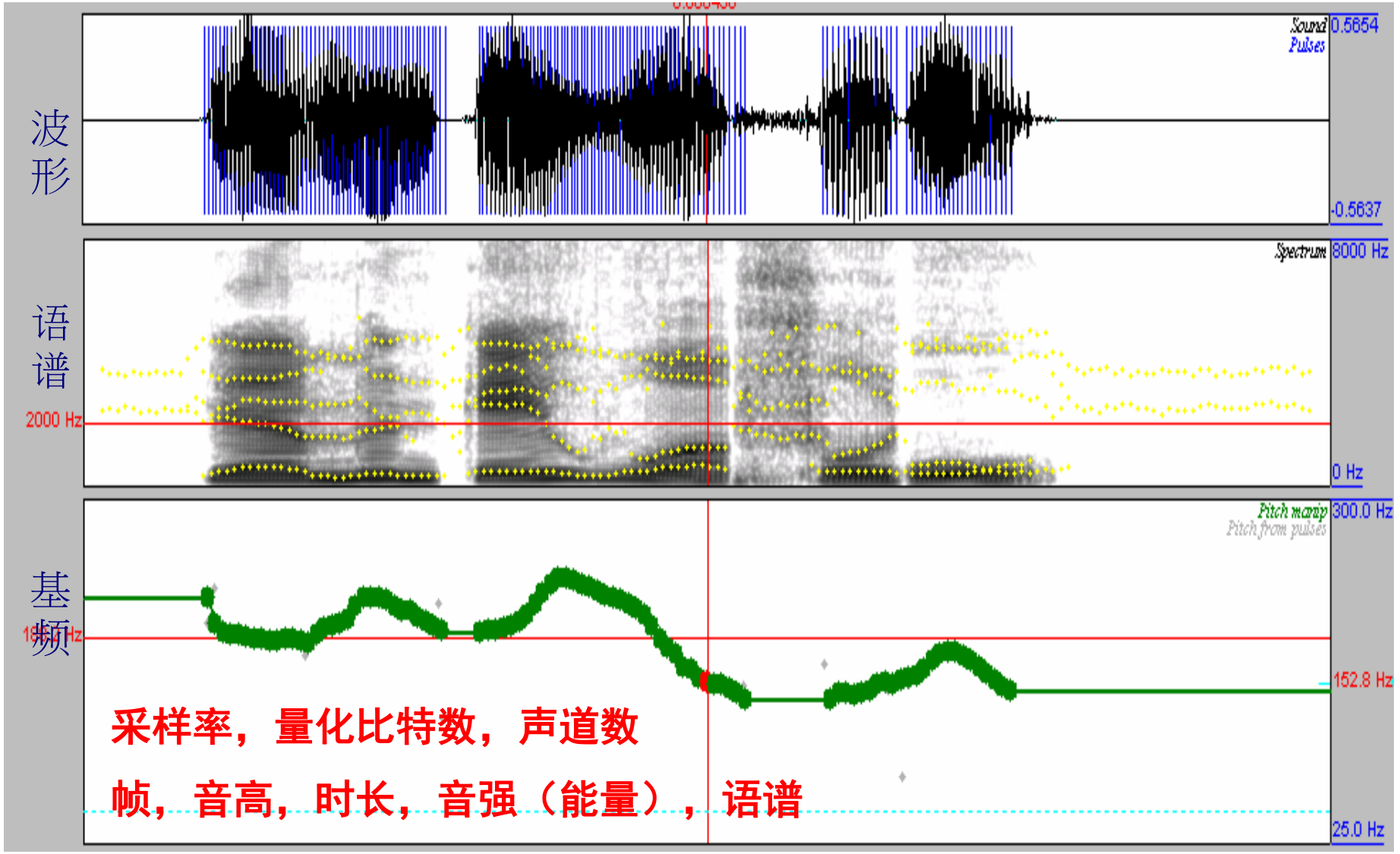


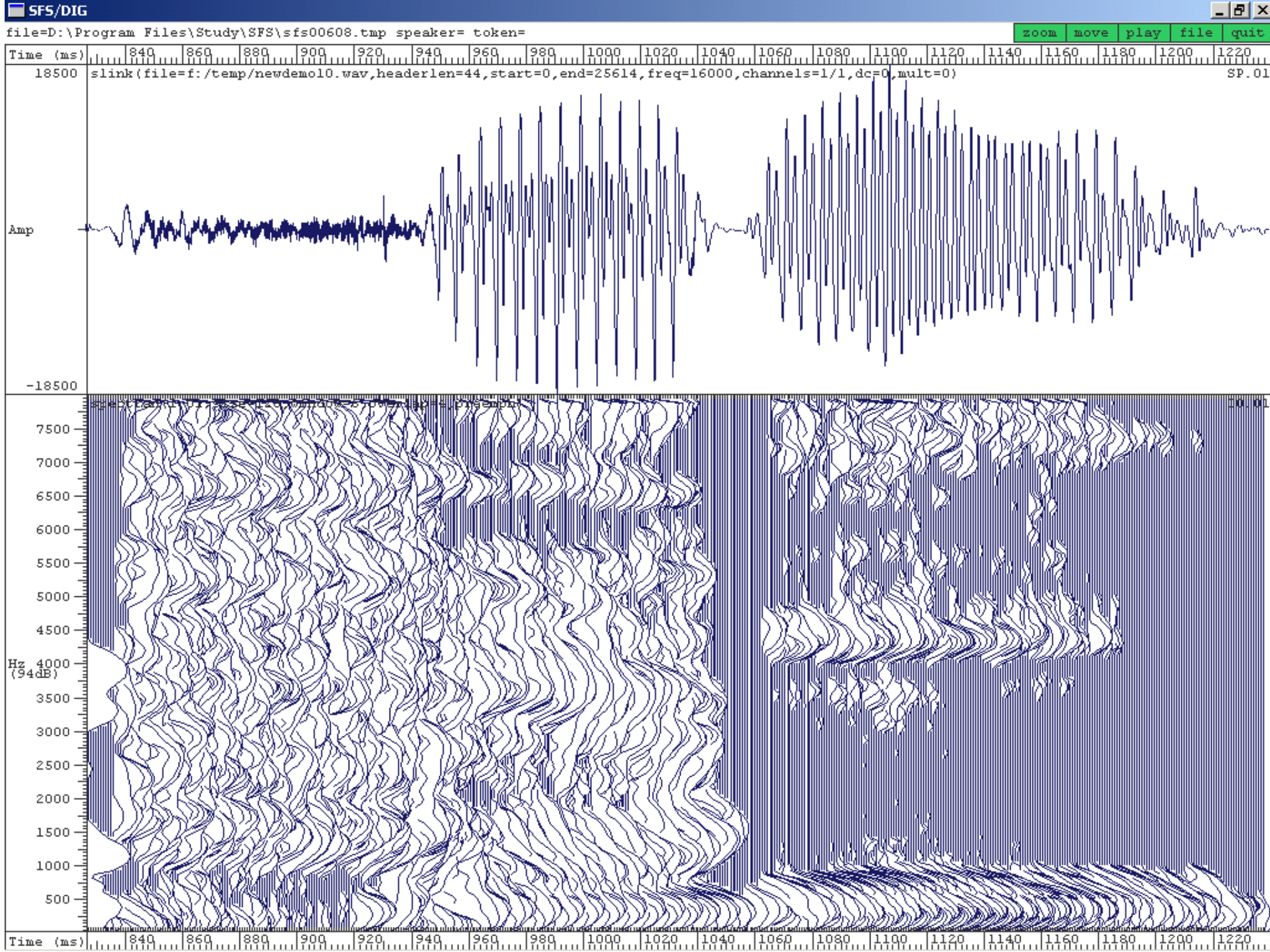
图 2-4　典型的声门脉冲串波形

语音在计算机中如何记录?



语音信号中的一些基本概念




SFS/DIG
file=D:\Program Files\Study\SFS\sfs00608.tmp speaker= token=
zoom
move
play
file
quit
Time (ms)
840
860
880
900
920
940
960
980
1000
1020
1040
1060
1080
1100
1120
1140
1160
1180
1200
1220
18500
slink(file=f:/temp/newdemo10.wav,headerlen=44,start=0,end=25614,freq=16000,channels=1/1,dc=0,mult=0)
SP.01
Amp
-18500
7500
7000
6500
6000
5500
5000
4500
Hz 4000
(94dB)
3500
3000
2500
2000
1500
1000
500
Time (ms)

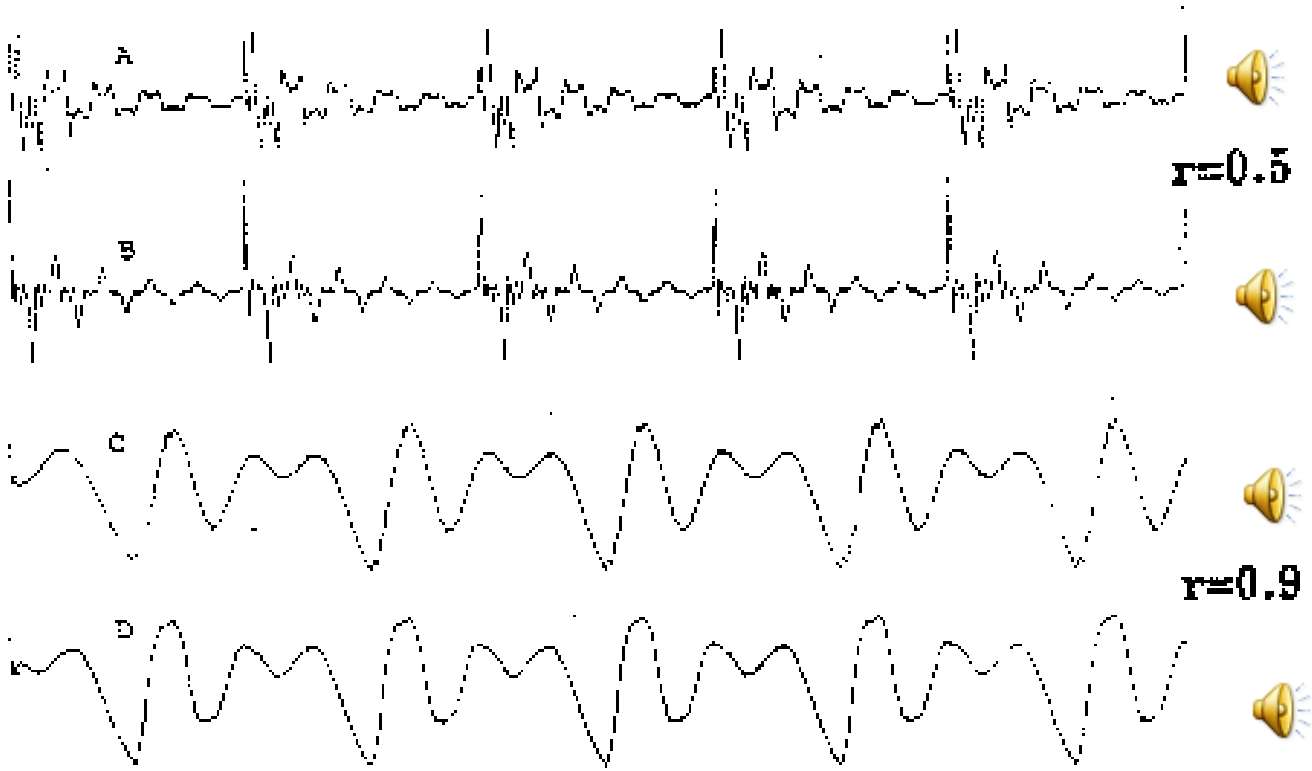
语音的特性1：波形不说明内容

[ə] a-b+c+d

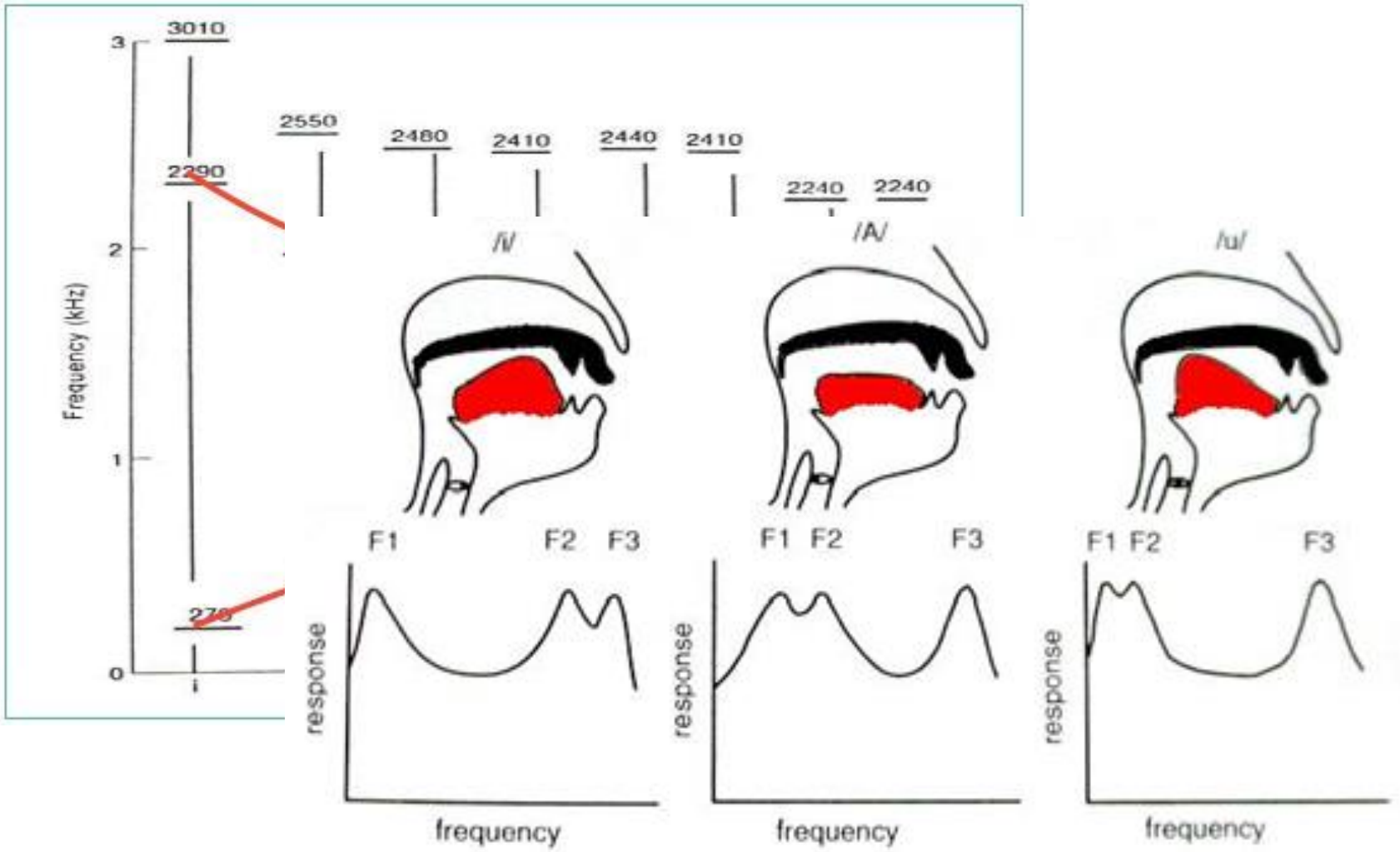
[ə] a+b+c+d

[i]

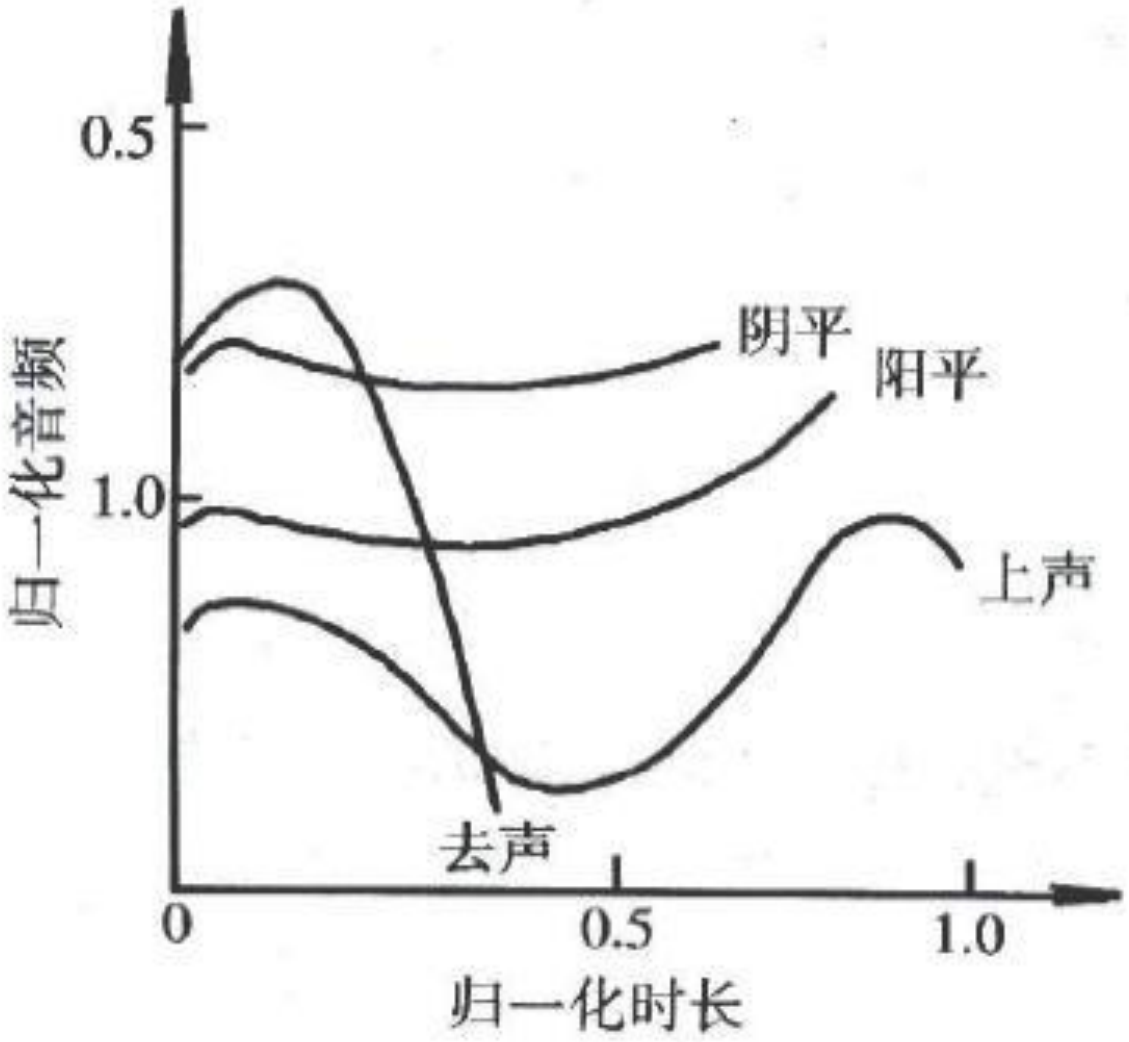
[u]



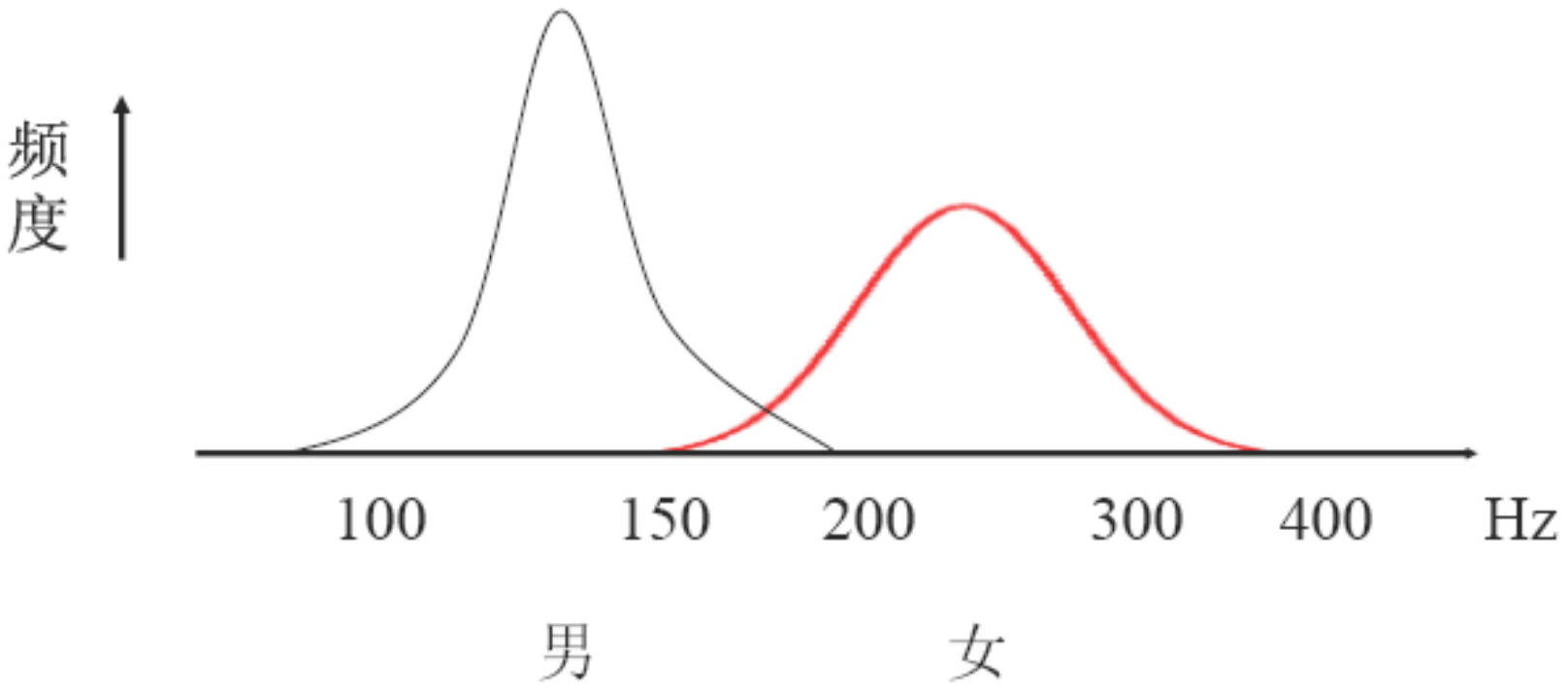
语音的特性2：共振峰基本决定内容



语音的特性3：基频决定声调



语音的特性4：声音一样又不同



- 不同发音人基频在对数频率轴上呈正态分布
- 男声均值125Hz，标准偏差20.5 Hz，女生分别为男生两倍

语音的特性4：汉语方言一级元音格局

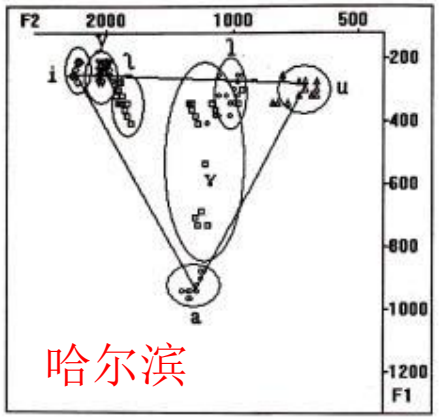


图 3.3 哈尔滨话一级元音声位图

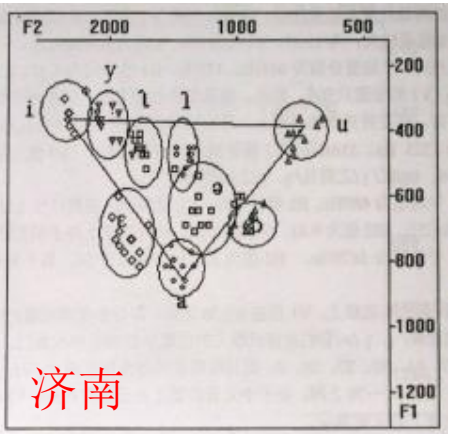


图 3.9 济南话一级元音声位图

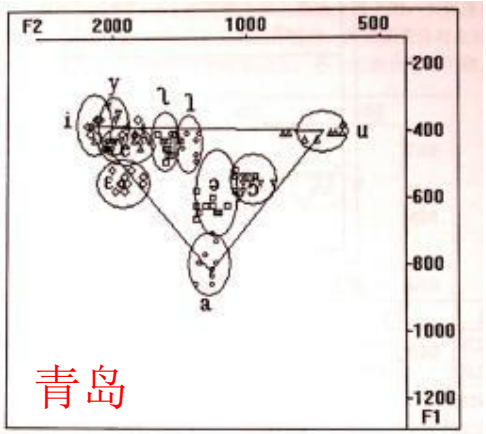


图 3.5 青岛话一级元音声位图

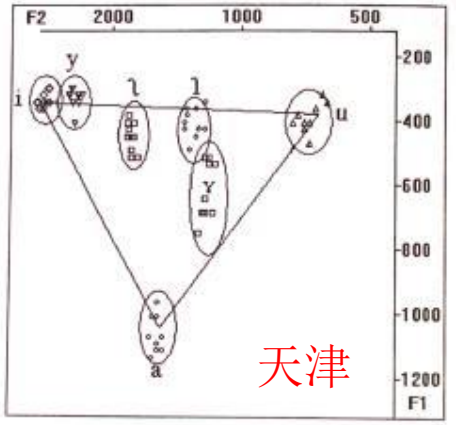


图 3.7 天津话一级元音声位图

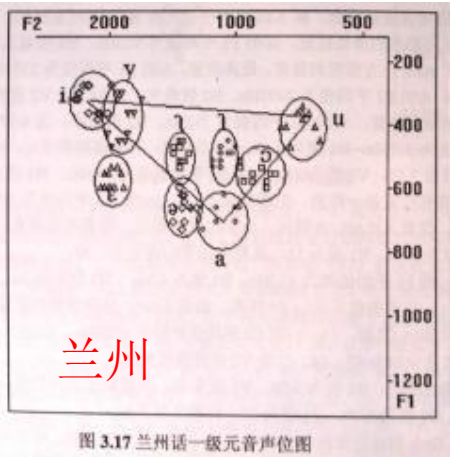


图 3.17 兰州话一级元音声位图

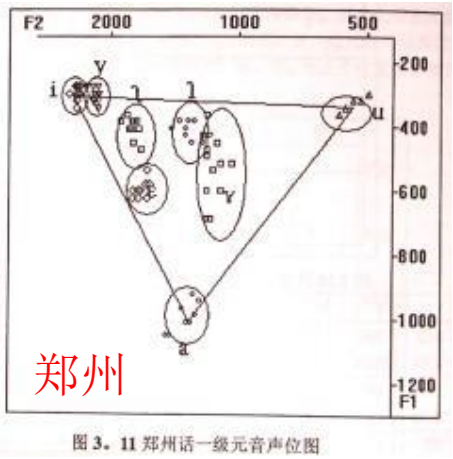


图 3．11 郑州话一级元音声位图

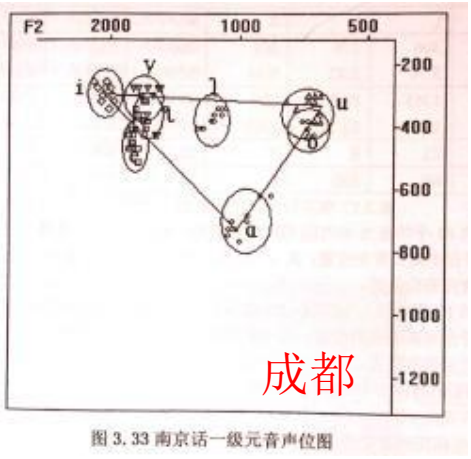


图 3. 33 南京话一级元音声位图

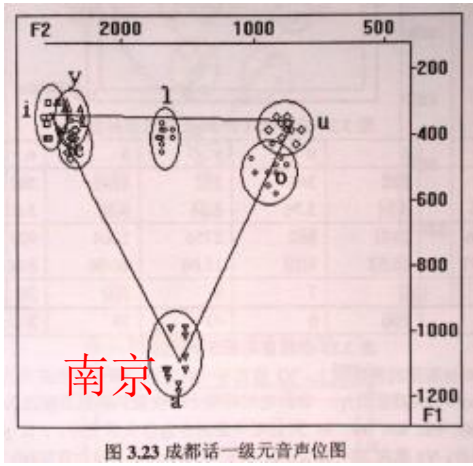
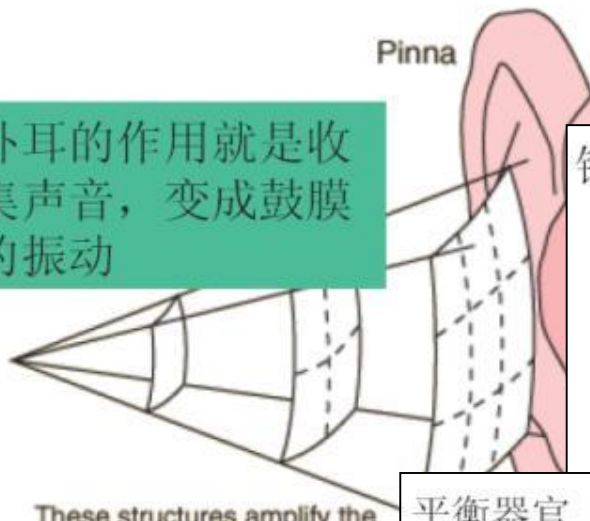


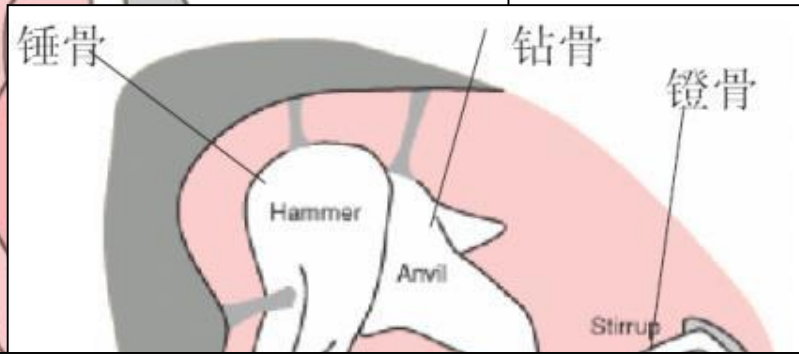
图 3.23 成都话一级元音声位图

引自：时秀娟，汉语方言元音格局的实验研究，南开大学博士论文，2005年4月。

人类听觉系统

外耳的作用就是收集声音，变成鼓膜的振动



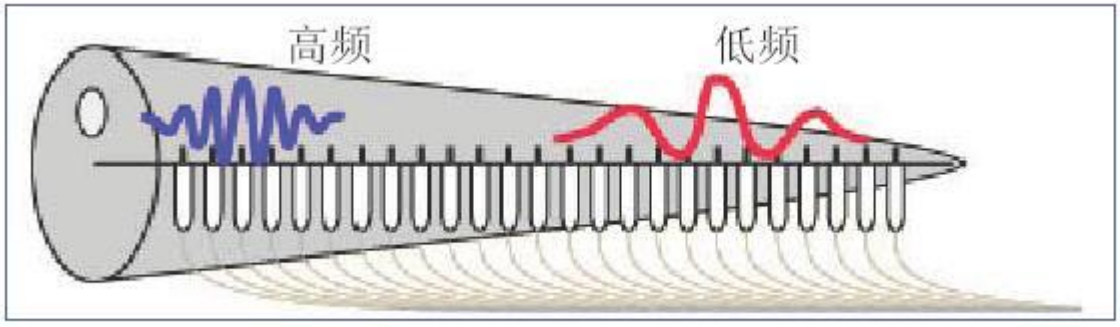


- 中耳由三小骨组成，连接鼓膜与卵圆窗
- 相当于传感器作用，将鼓膜上的机械振动放大到卵圆窗膜的振动（3倍）

中内淋

，防止耳的损

- 内耳由半规管和耳蜗组成
- 半规管是人体平衡器官，和听觉无关
- 耳蜗是主要的听觉部分，相当于人体麦克风，将来自外耳的声波压力信号转换成电脉冲，并经听神经传送到大脑



人体麦克风

计算机眼中的语音

- 语音必须转成数字信号，计算机才能处理
- 通常需要在语音的数字信号中，提取出各种参数，如：基频、频谱（MFCC）、时长等。
- 运用一系列统计模型或神经网络模型进行语音建模

语音和语言从来都不分家

- 语言：人类特有的能力
- 有2500至3500种语言
- 汉语属汉藏语系，英语属印欧语系日耳曼语族
- 语言层级：
  - 音素、声韵母、音节、字、词、短语、句子、篇章
  - 以有限的音节和字按规定的文法构建出无限的句子
- 计算机必须要将语言信息和语音信息结合起来，才能实现语音识别、语音合成等技术。缺乏语言信息的支持，计算机只能停留在语音的信号层面的处理。

计算机眼中的语言

“语音合成技术十分先进了”

在分词之前，计算机看到的结果是：

D3 EF D2 F4 BA CF B3 C9 BC BC CA F5 CA AE B7 D6 CF C8 BD FB C1 CB A1 A3 00

分词过后，计算机看到的是：

名词 动词 名词 副词 形容词 助词 标点

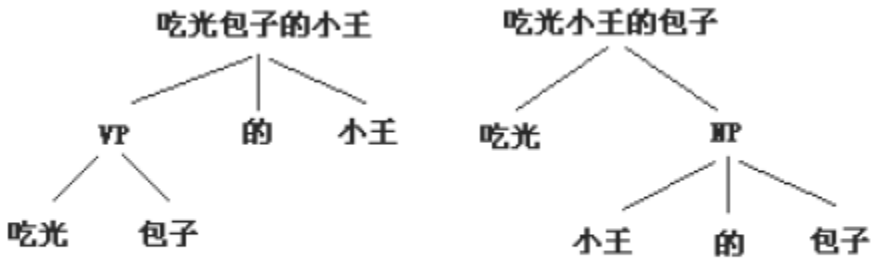


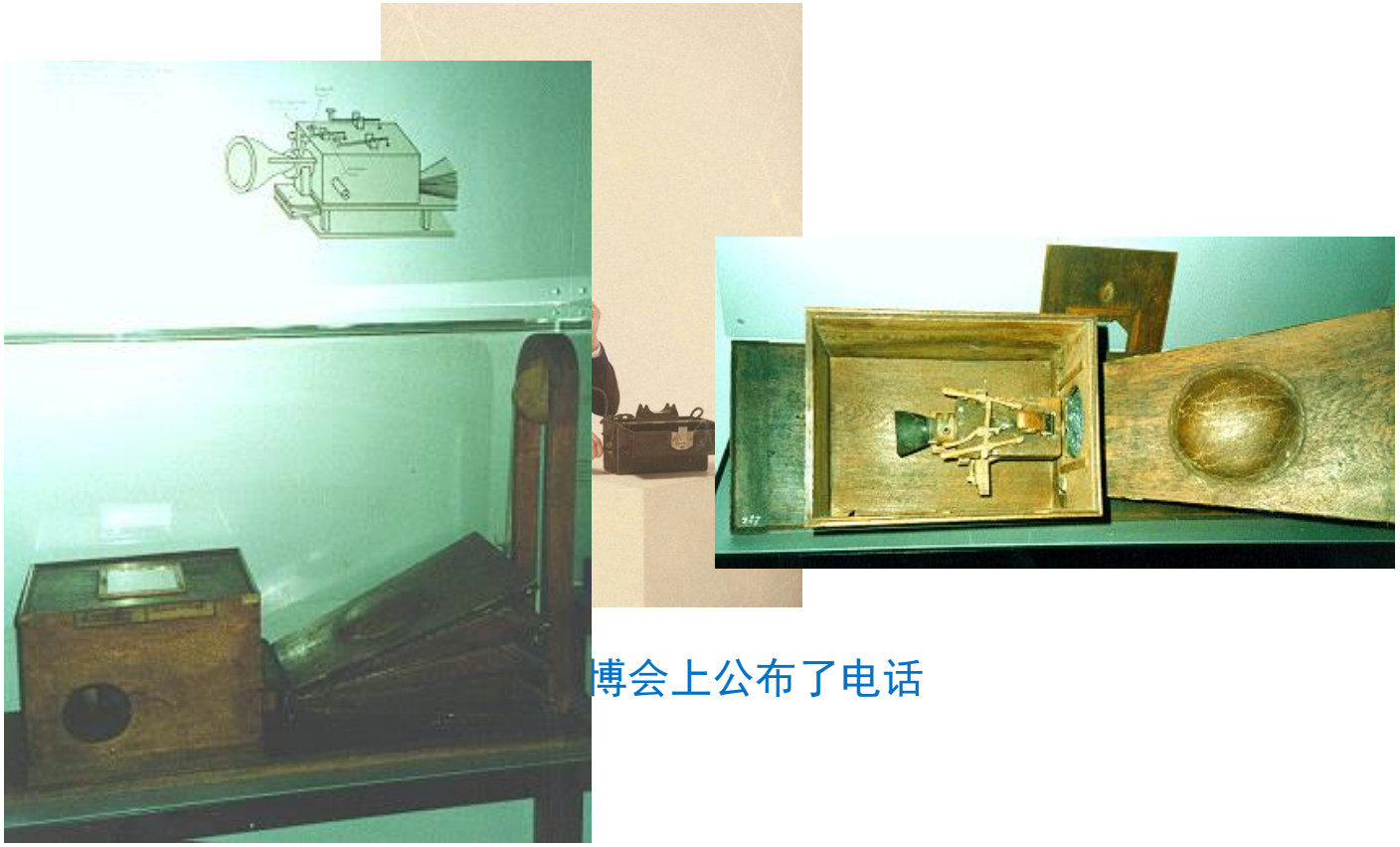
图 1：相同词类序列不同拼接结果的示例

- 数据少意思多！寥寥数字，无穷意境

本节课提纲

- 语音基本概念
- 语音研究历史
- 语音技术概述

语音研究历史



博会上公布了电话

贝尔发明电话

1876

60年代以前

起步阶段

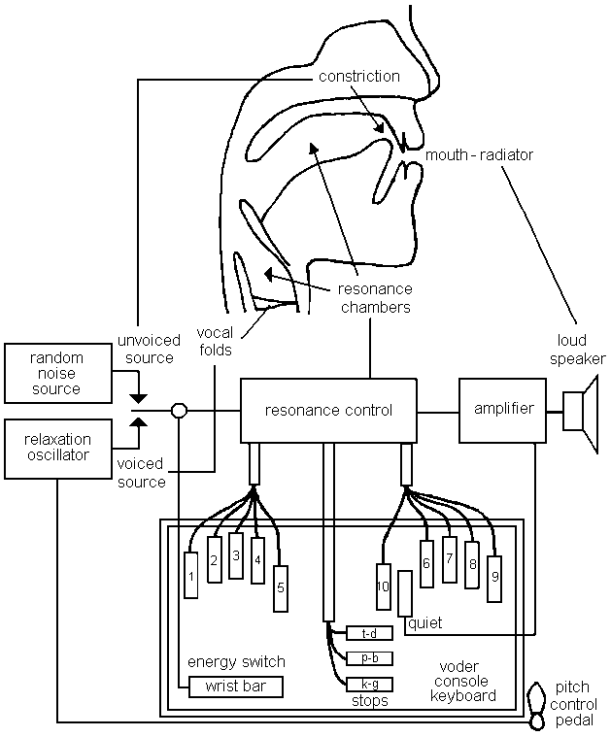
60年代

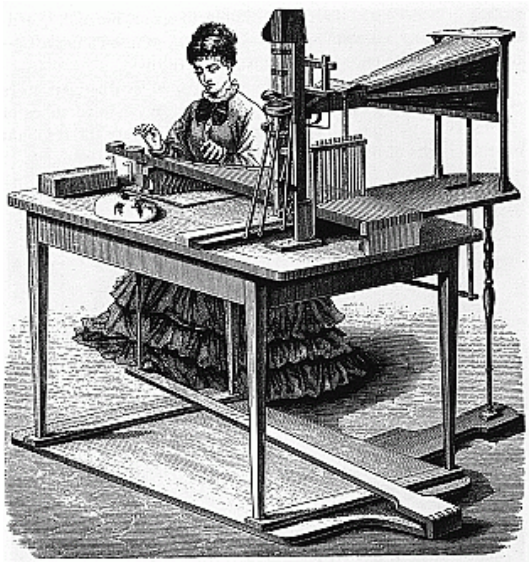
70年代

80年代至2010年

2010年后

语音研究历史





H.Dudley研制成声码器

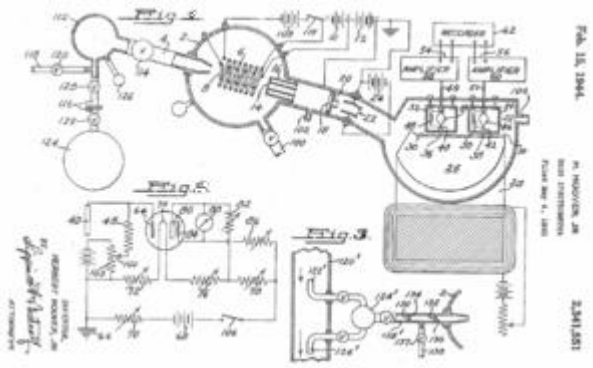
1939年H. Dudley研制成功第一个声码器

贝尔发明电话

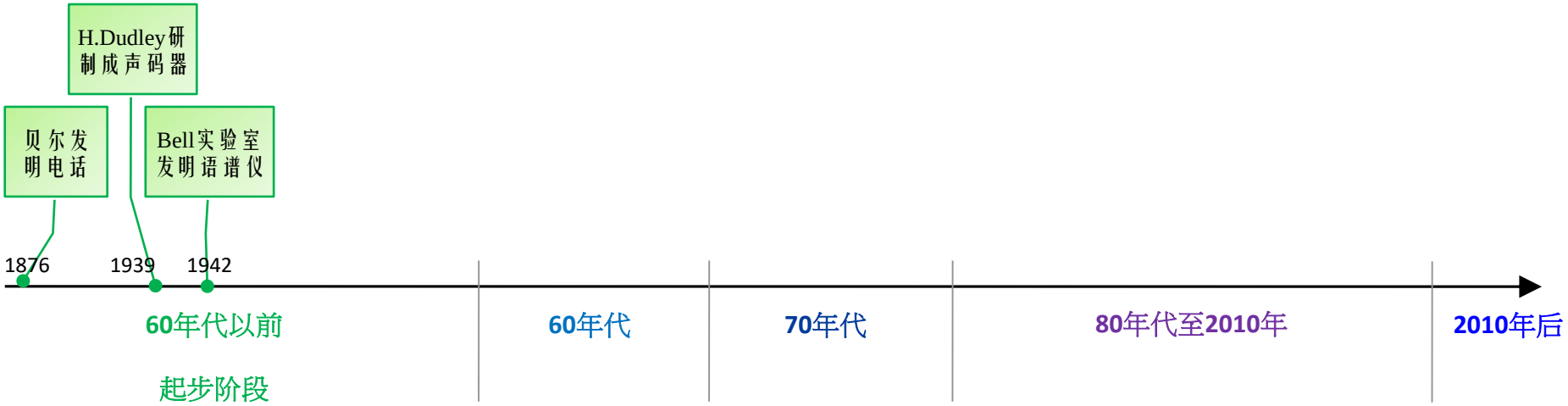
1876 1939

60年代以前 起步阶段 | 60年代 | 70年代 | 80年代至2010年 | 2010年后

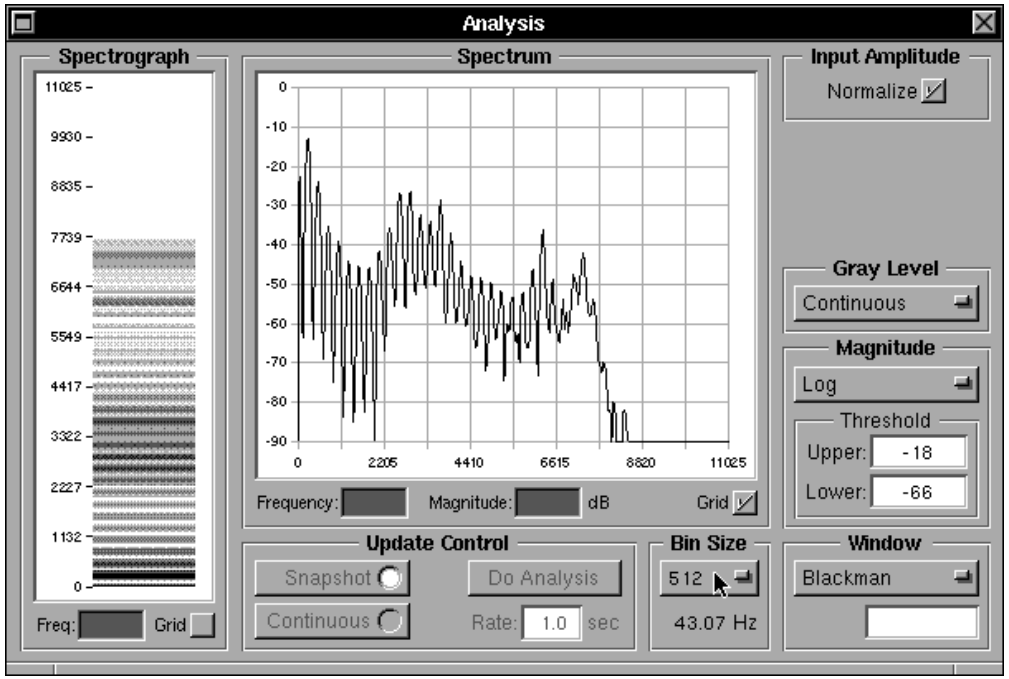
语音研究历史



1942年Bell实验室发明了语谱仪

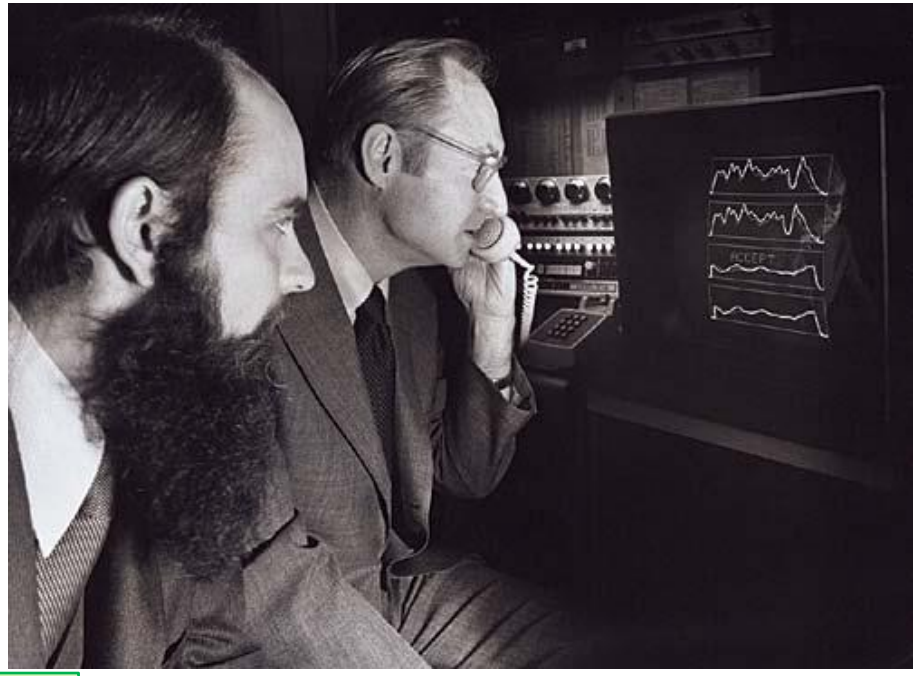


语音研究历史



1948年美国Haskin实验室研制成功“语图回放机”

语音研究历史



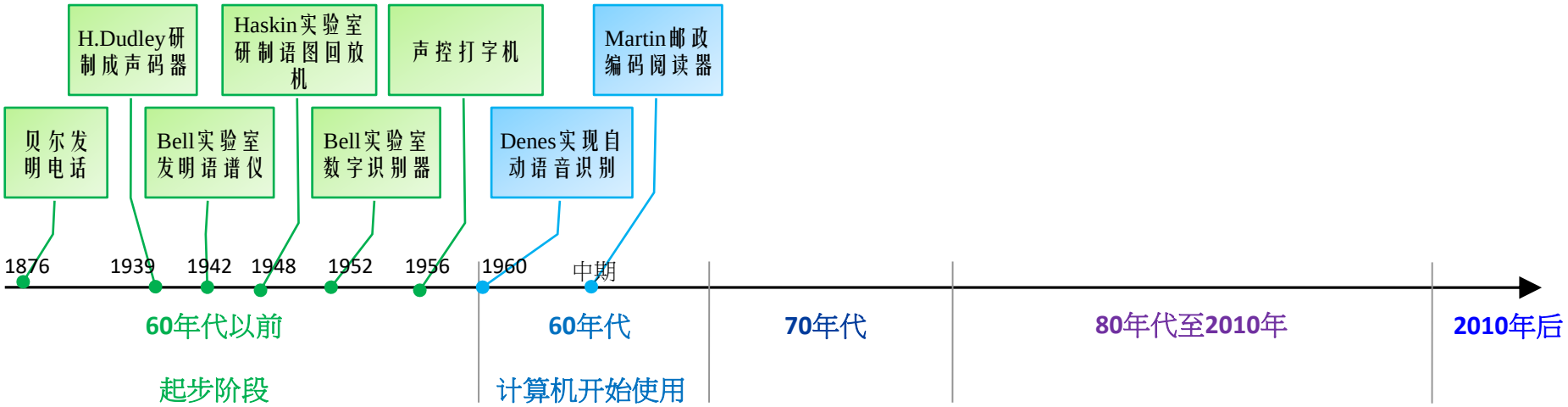
l l实验室研制成能识别十个英语数字的识别器

语音研究历史

•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。

•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。

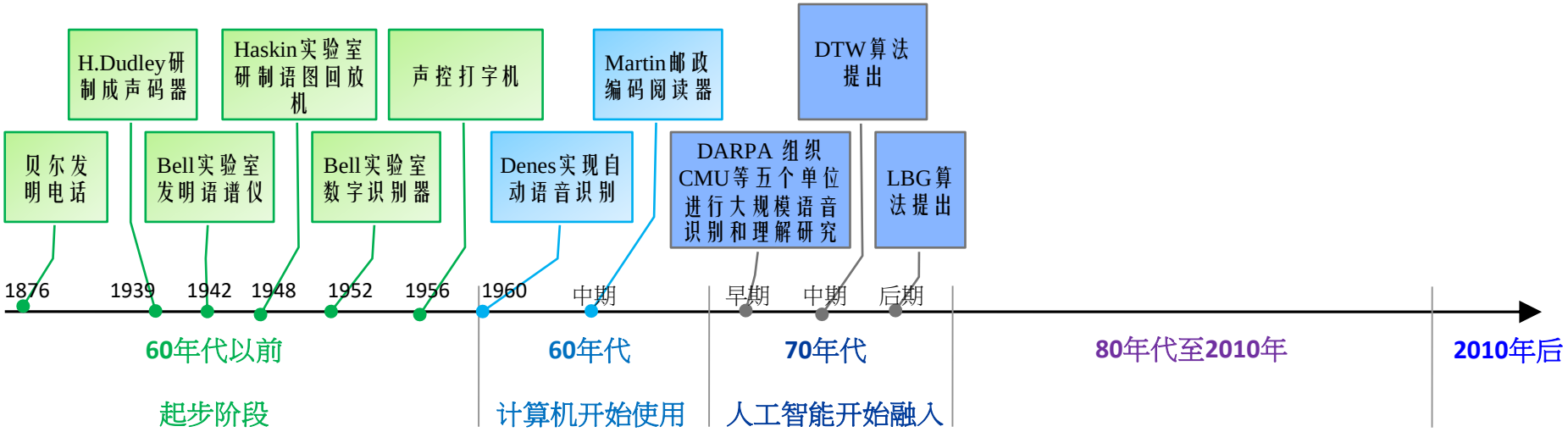


语音研究历史

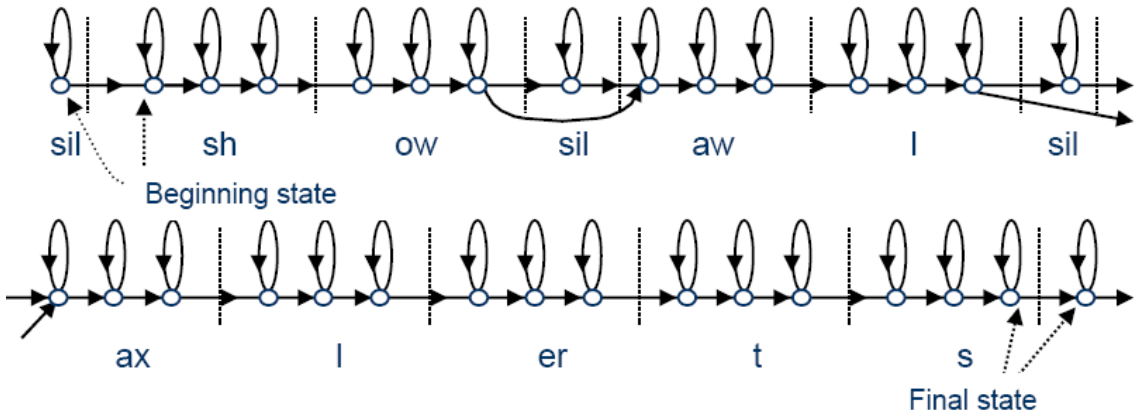
- 70年代开始，人工智能技术开始引入到语音识别中。美国国防部ARPA组织了有CMU等五个单位参加的一项大规模语音识别和理解研究计划。

- 70年代中，日本学者提出的动态时间弯折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。

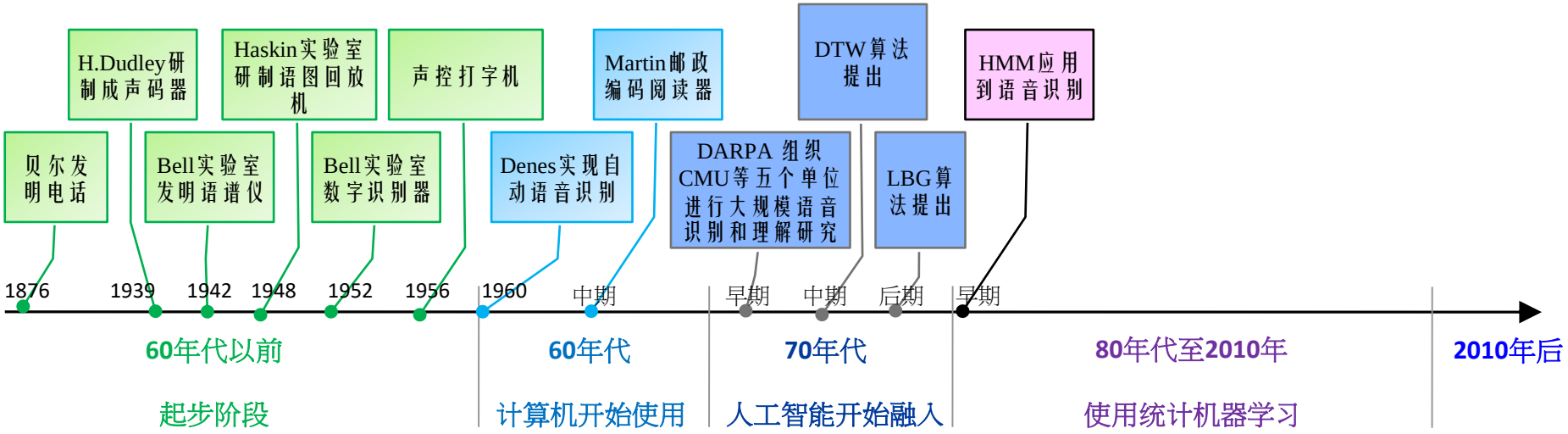
- 70年代末，基于矢量量化码本生成的LBG算法被提出，从而使矢量量化技术广泛地应用于语音识别、语音编码和说话人识别中。



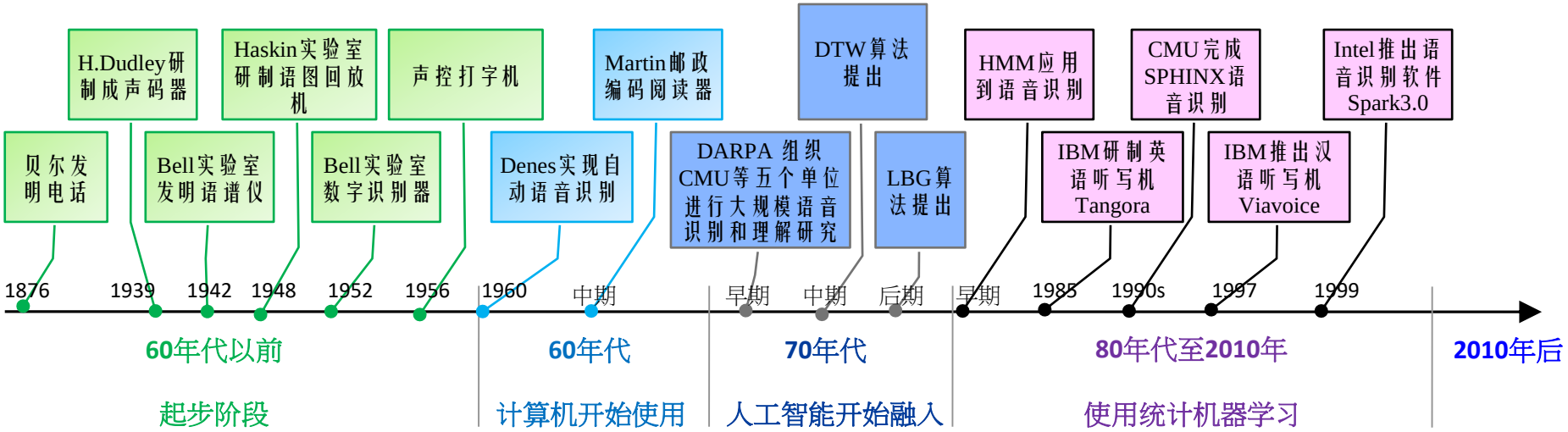
语音研究历史



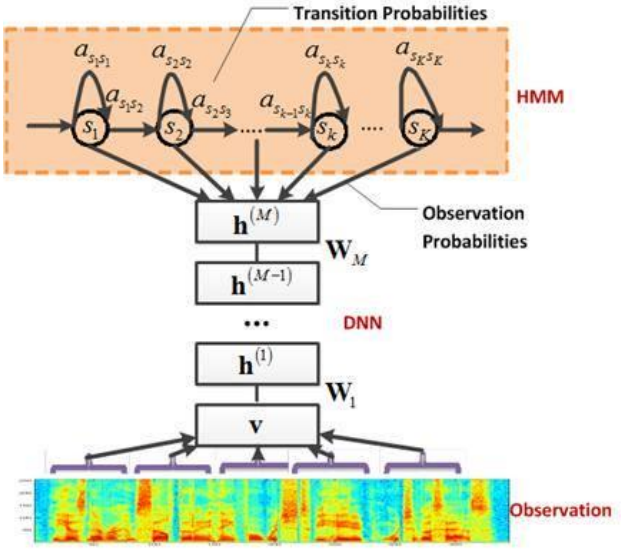
从70年代末80年代初开始，HMM 技术被应用到语音识别中



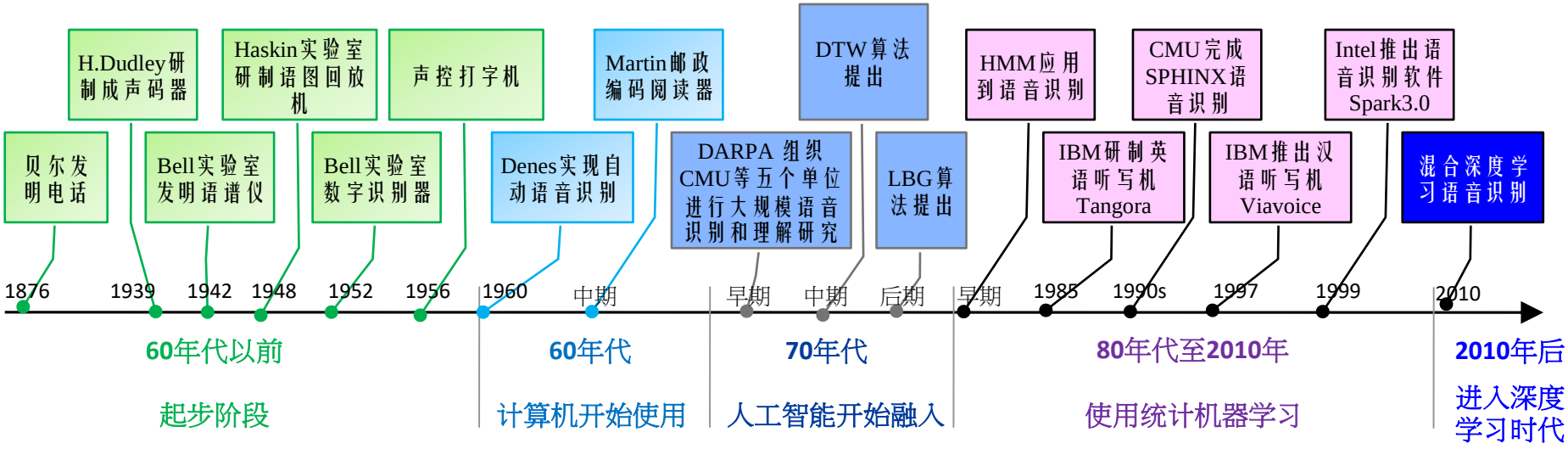
语音研究历史


H.Dudley研制成声码器
Haskin实验室研制语图回放机
声控打字机
Martin邮政编码阅读器
DTW算法提出
HMM应用到语音识别
CMU完成SPHINX语音识别
Intel推出语音识别软件Spark3.0
贝尔发明电话
Bell实验室发明语谱仪
Bell实验室数字识别器
Denes实现自动语音识别
DARPA 组织CMU等五个单位进行大规模语音识别和理解研究
LBG算法提出
IBM研制英语听写机Tangora
IBM推出汉语听写机Viavoice
1876
1939
1942
1948
1952
1956
1960
中期
早期
中期
后期
早期
1985
1990s
1997
1999
60年代以前
起步阶段
60年代
计算机开始使用
70年代
人工智能开始融入
80年代至2010年
使用统计机器学习
2010年后

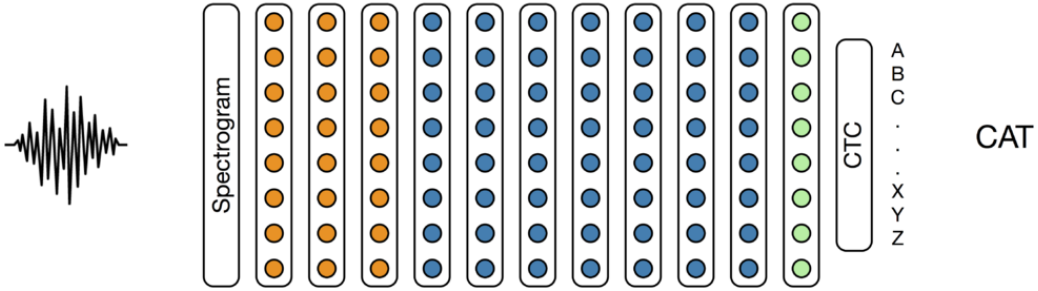
语音研究历史



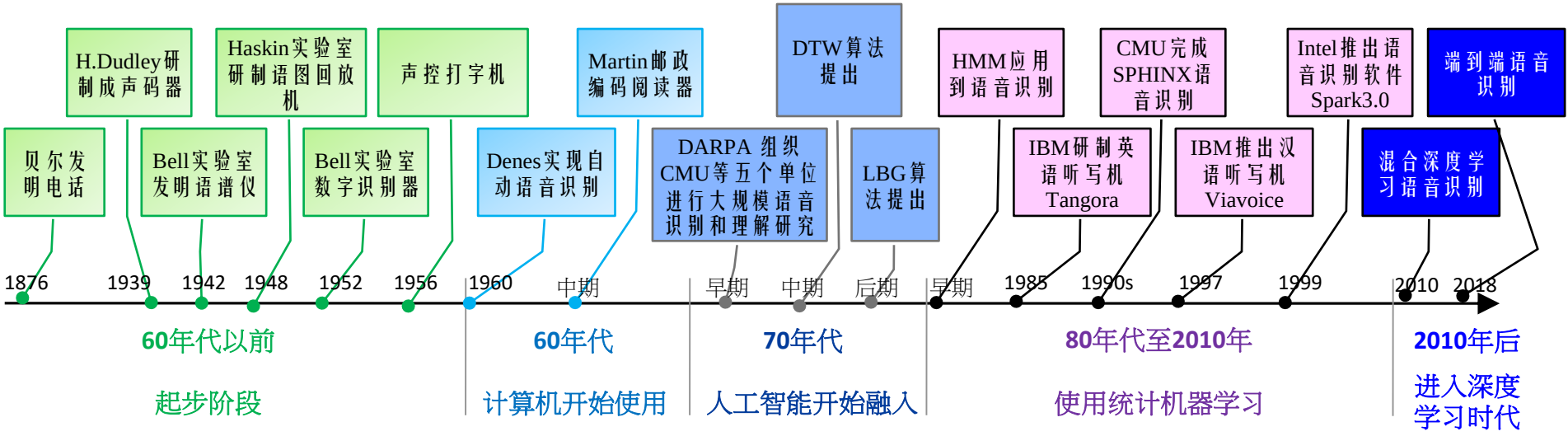
混合深度学习语音识别



语音研究历史



端到端语音识别



本节课提纲

- 语音基本概念
- 语音研究历史
- 语音技术概述

语音识别（ASR）

- **语音识别（ASR）： 把声音变成文字(耳朵的功能)，相当于给机器装上了人工的耳朵。**

孤立词识别技术

连续语音识别

关键词识别技术

话者识别技术

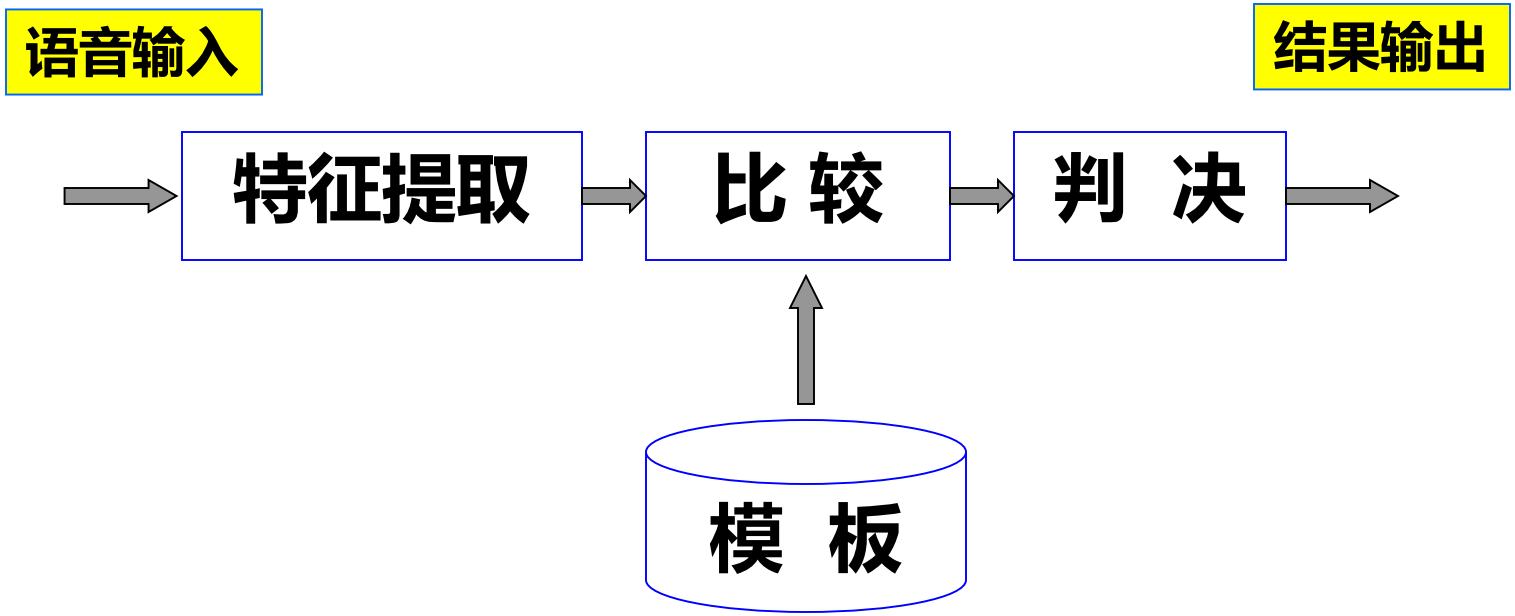
语音识别的复杂性

语音识别的复杂性

- **孤立词/连续语音？** Isolated or Continuous speech
- **认人/不认人？** Speaker-dependent or Independent
- **小词汇量/大词汇量？** Small or large vocabulary
- **安静环境/噪杂环境？** Environment robustness
- **一般信道/电话信道？** Channel adaptability

语音识别技术

最基本的孤立字识别系统



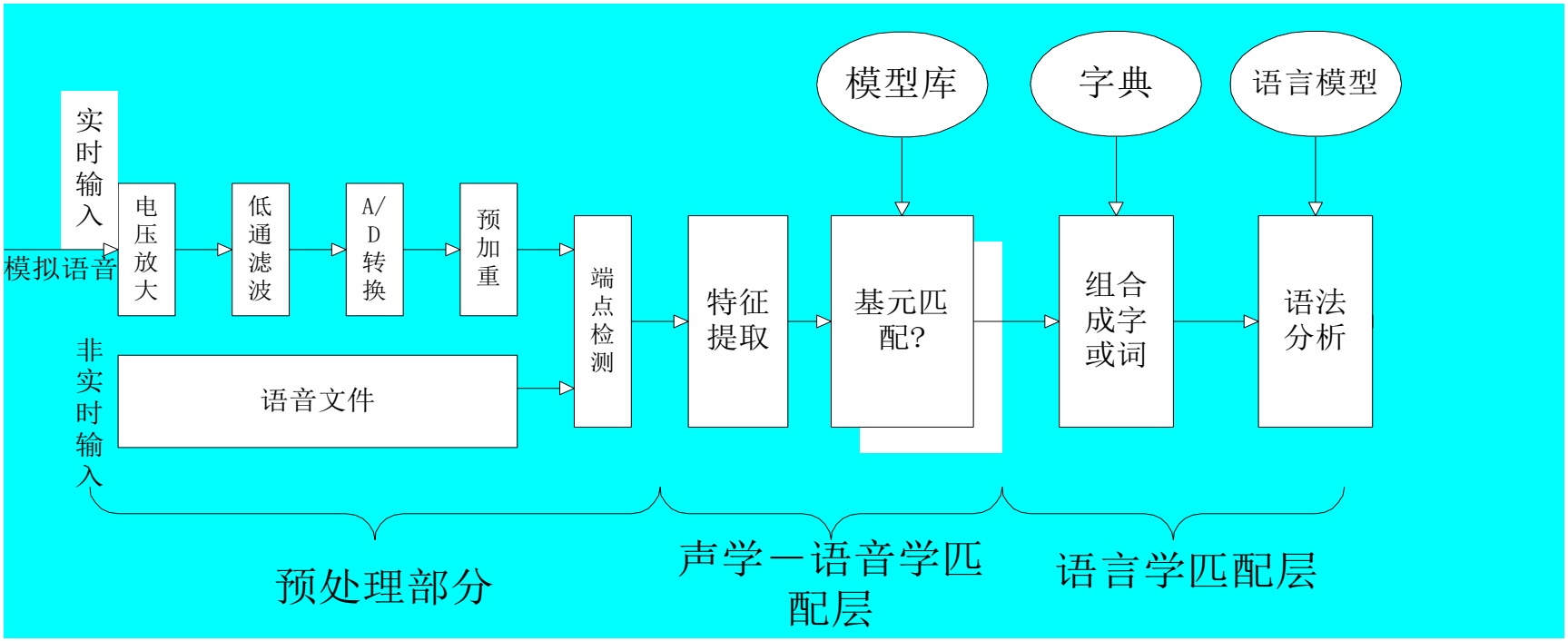
语音识别技术

基于统计的语音识别系统组成

- **前端处理（特征参数提取）**
  - **最大限度地冗余信息的剔出， 和最大限度地语音的区别特征的保留。例：**LPC，LSP，DFT，MFCC。
- **模型的建立与学习（生学模型、语言模型）**
  - **声学模型建立与学习：模板，HMM。**
  - **语言模型建立与学习：词 BI-GRAM，TRI-GRAM,POS BI-GRAM，**
  - **有监督学习和无监督学习**
  - **自适应学习：OFF LINE 有监督与无监督，ON LINE 无监督**
- **识别（分类）**
  - **最佳路径搜索，决策最可能的结果**
  - **ROBUST性**

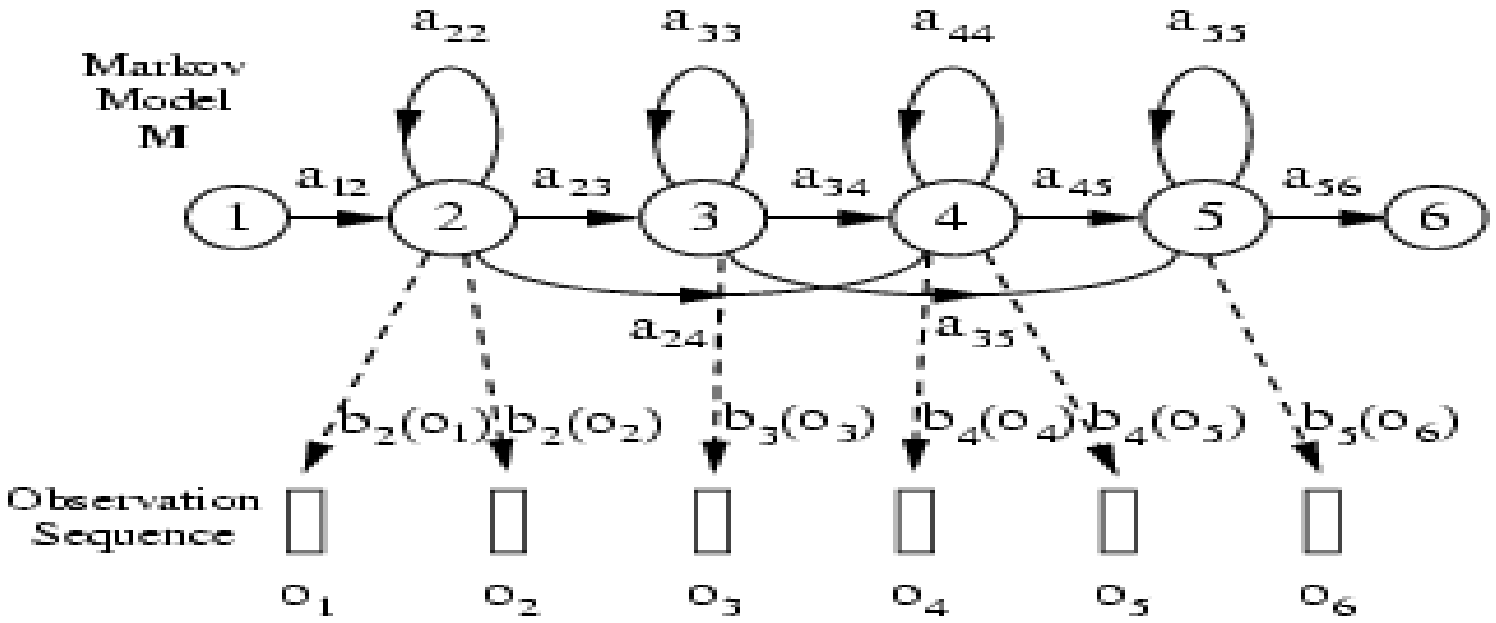
语音识别技术

基于统计的语音识别系统系统构成图

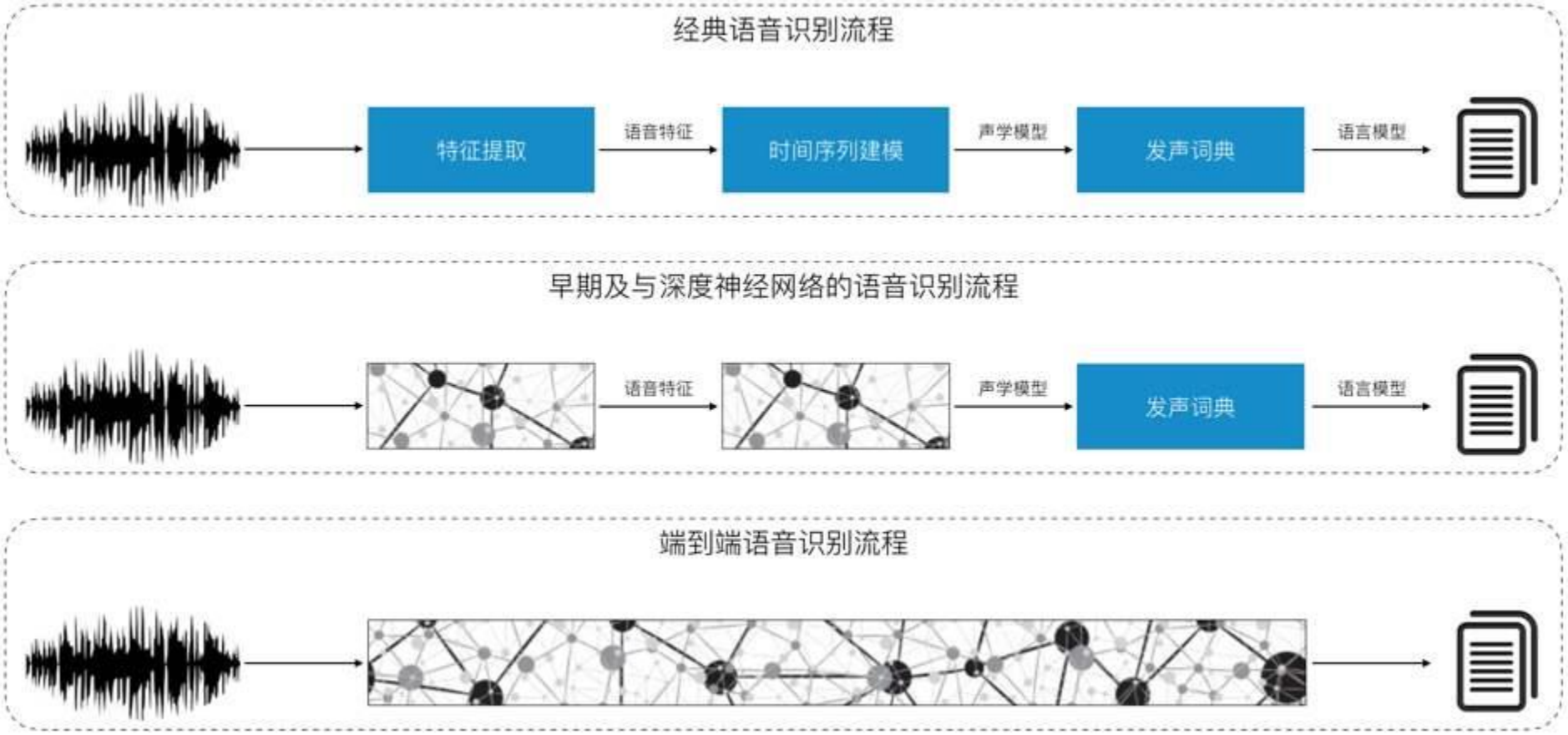


语音识别技术

- **隐马尔可夫模型（HMM）**



从统计机器学习到深度学习



电话语音识别技术

电话语音识别技术的特点

- 电话信道环境下的非特定人连续语音识别
- 考虑到电话信道特性，噪音，话机的差别等因素的语音识别ROBUST问题的研究
- 电话信道环境下的非特定人连续语音识别数据库的建立

现有ASR的技术应用

- 近年来ASR核心研究的前进步伐放慢，性能几乎饱和，以云计算为基础的语音识别获得大量应用
- 但是现有系统还存在很多问题
  - 使用时经常需要用户很好配合
  - 在复杂场景下，识别性能下降明显，现有的信号处理方法收效甚微
  - 面对口语化严重、对话中出现不符合语法的病句、集外词、任务外的词、说话习惯的嗯啊....等，现有的系统难以胜任
  - 混合语言依然没有很好解决
- 这样的识别系统只要用户界面设计、实现的好，可以发挥其应有的价值！

身份识别和确认（声纹识别 ）

- 功能：通过语音识别或确认说话人身份
- 分类：
  - 身份确认、身份识别
  - 文本相关、文本无关
- 难点：
  - 相同人不同身体状态的音色有差别
  - 要防止恶意的模仿（DeepFake已经成为一个研究热点）
- 方法：
  - GMM，HMM、iVector、DNN
- 水平：
  - 1000个人，97%以上的识别正确率

语音增强

- 功能：将语音从噪声中分离出来
- 难点：
  - 某些噪声很像语音；
  - 有些语音也算噪声；
  - 降噪效率
- 方法：
  - 对语音和噪声分别建模
  - 噪音快速建模算法

语音合成

- **语音合成（TTS）：主要解决的问题是将文本状态的文字信息转化为可听的声音信息。**

语音问答系统

自动阅读

信息查询

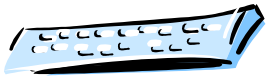
语言学习系统

语音合成的历史

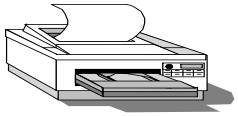
- 电子计算机发明以后，语音合成技术得到了飞速的发展，方法也发生了根本性的变化

年代	里程碑
1939	**Bell**实验室发明**voder**，最早的现代合成语音产生方式
1960's	共振峰参数化合成器结合规则合成
1986	**TD-PSOLA**算法发明，显著提高了合成音质
1990's	**ATR**提出大语料库合成方法，使语音合成最终达到市场实用化效果
2000-2010	**HTS**系统获得大规模应用，语音合成的自然度获得很大提升
2011-	深度学习语音合成获得大规模应用，声音在音质、韵律表现力逼近真实人的声音

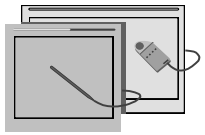
语音合成技术



键 盘



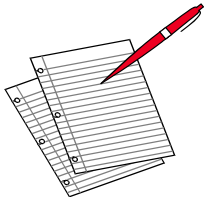
光电扫描



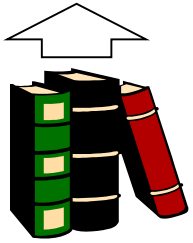
手写识别

网络/数据库

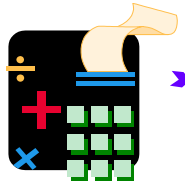
文 本



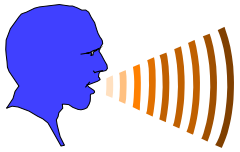
语言处理



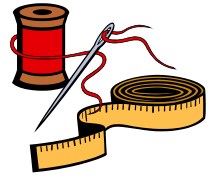
词典/规则



韵律处理



结果输出



语音合成

语音库

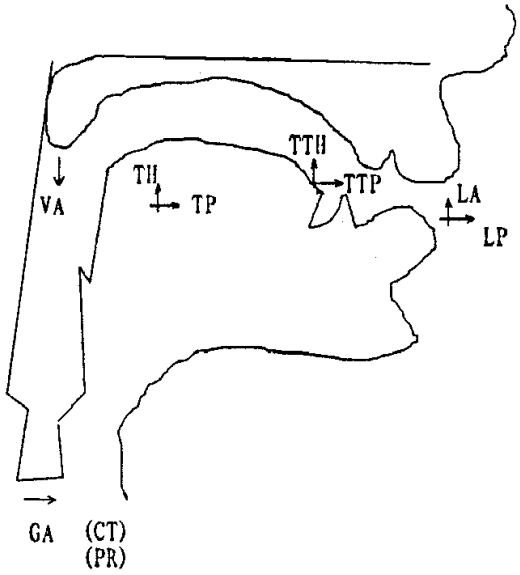
语音合成技术

语言合成技术之一 Articulatory Synthesis

- 根据人类发音机理方式工作的合成方法
- 模型主要组成部分
  - 声门波发生装置 --- 声带
  - 气管-口腔声道腔体模型
  - 嘴唇的辐射模型
- 主要优缺点
  - 真实的反映了人类发音的整个过程
  - 人类发音过程的模型不够精确，无法得到清晰度高的语音

语音合成技术

- **语言合成技术之二 Source-filter Synthesis**
  - **基于语语音数据信号处理的合成方法**
  - **模型主要组成部分**
    - **声门波激励源**
    - **描述声道模型的滤波器**
  - **主要优缺点**
    - **合成语音的音质比上一种方法有很大的提高，但是仍然不是很好**
    - **可以对合成语音在音色和声调上进行较为灵活的调整**

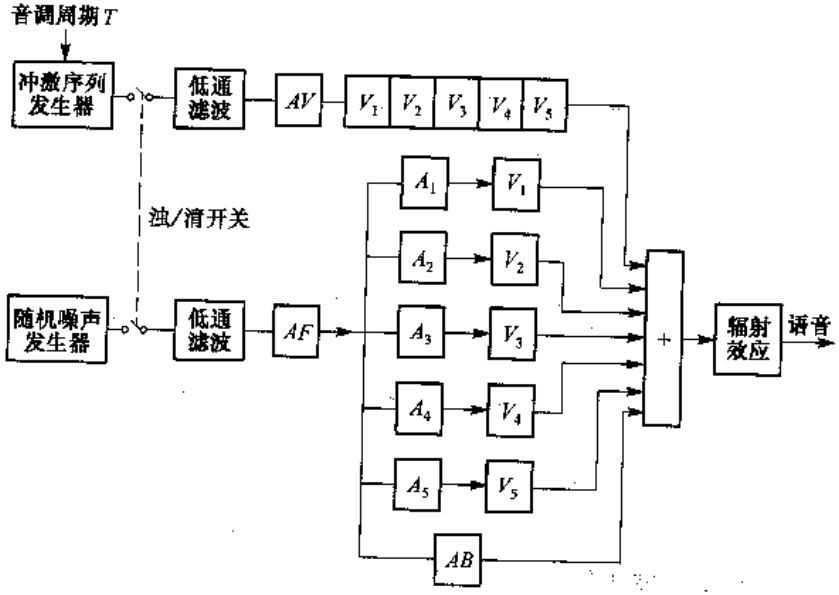


发音参数

LA	唇的开口度
LP	唇的突出度
TTH	舌尖高度
TTP	舌尖前后位置
TH	舌体高度
TP	舌体前后位置
VA	小舌位置

声源参数

GA	声门开度
CT	声带张力
PR	肺气压



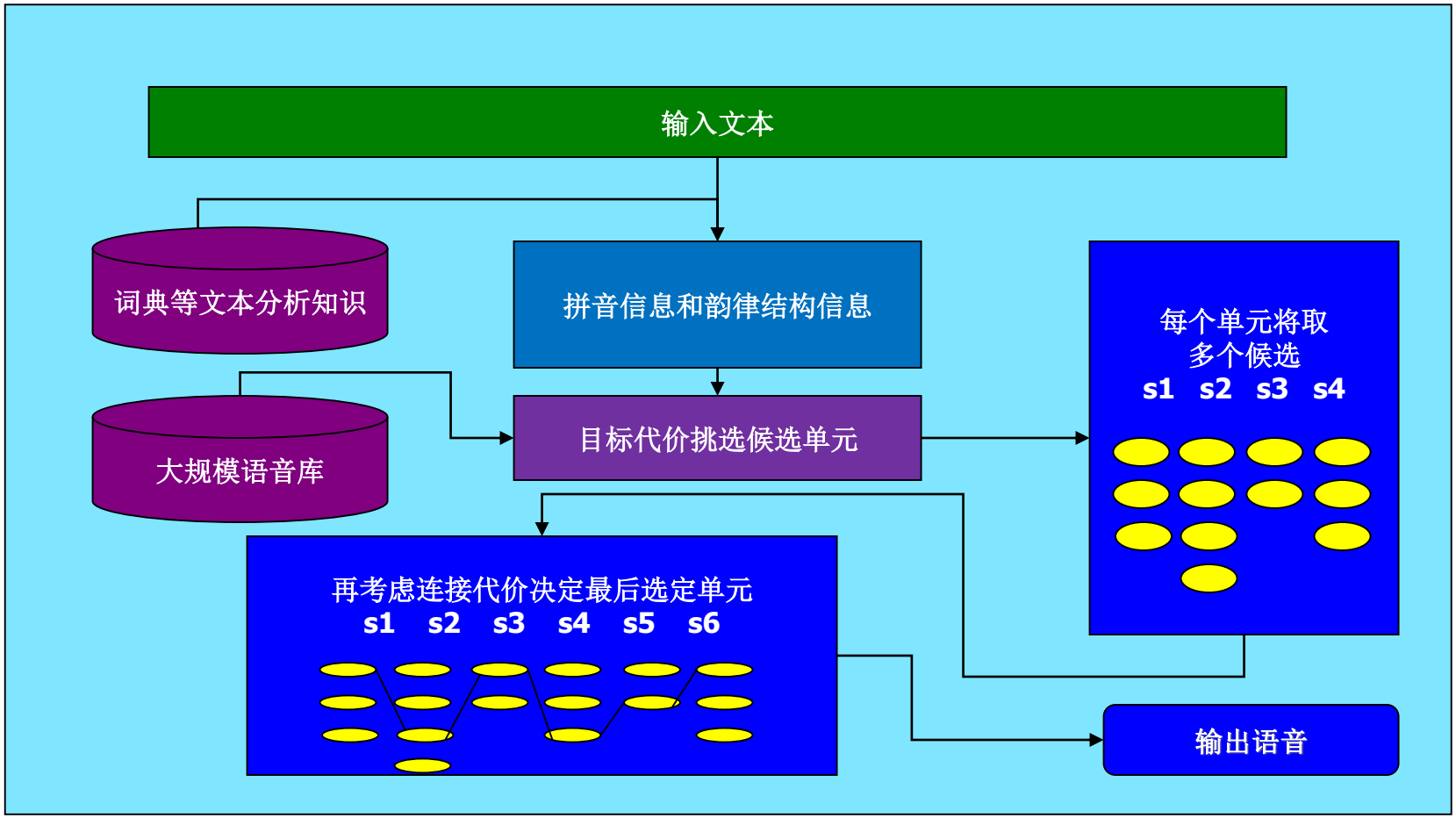
语音合成技术

语言合成技术之三 Concatenative Synthesis

- 利用原始语音片断作为合成单元
- 关键技术
  - 原始语音片断的获取方法
  - 原始语音片断的挑选方法和拼接算法
- 主要优缺点
  - 合成语音的音质比上两种方法有质的提高，因为不需要进行大的调整
  - 语料库的录制和制作工作量巨大，同时合成语音的灵活性较低

语音合成技术

拼接语音合成系统处理流程



语音合成技术

拼接语音合成系统效果

年份	1995年	1998年	1999年	2001年	2003年
自然度	<3.0	3.0	3.5	3.8	4.3




最新中文拼接系统

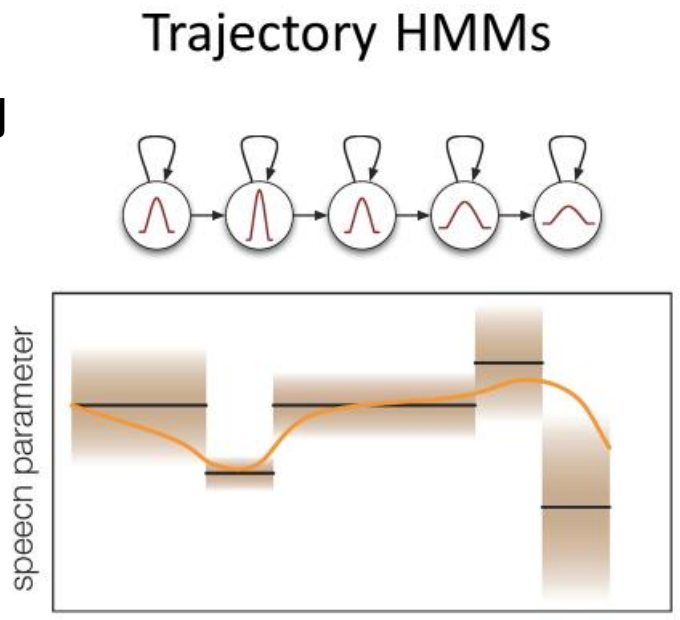
语音合成技术

语言合成技术之四 HMM Based TTS

- 利用HMM模型直接对语谱和韵律进行建模
- 关键技术
  - HMM模型对特定人的语音进行建模
  - 良好的合成器对预测出来的语谱和韵律参数进行合成
- 主要优缺点
  - 合成语句自然流畅，普适性好
  - 能够容易的模拟各种不同的说话人，不同情感，不同语气，但效果总体有限
  - 因为采用合成器进行语音的合成，清晰度难以提高

女声

男声



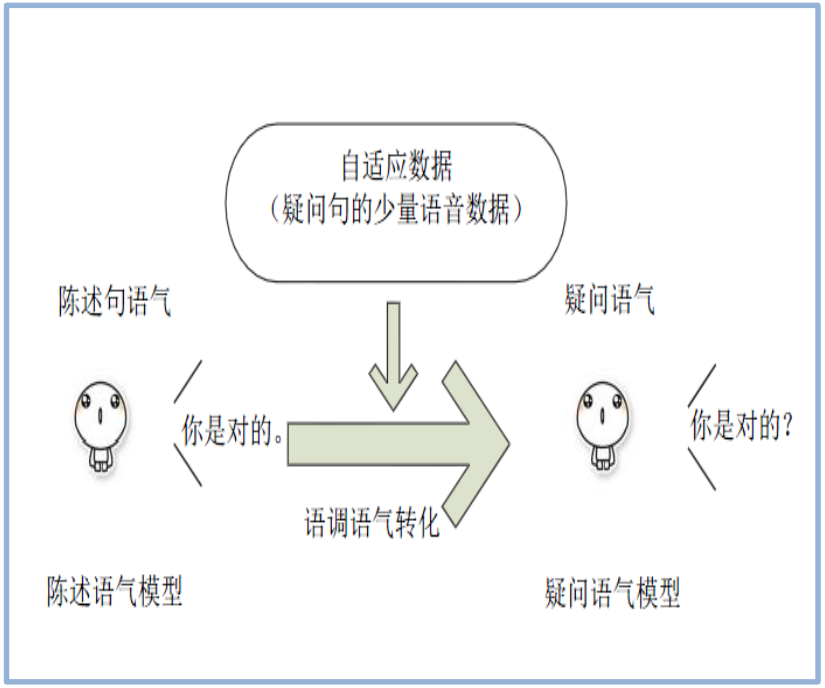
语音合成技术

- **语言合成技术之五 多模态语音合成**
  - **将语音合成与嘴唇、舌位、脸部运动结合起来**
  - **关键技术**
    - **对语音内容和嘴唇运动以及脸部运动进行同步**
    - **利用三维模型或者是图像录像进行脸部图像的生成**
  - **主要适用场合**
    - **可视聊天等各种沟通方式中（例如于msn，QQ等结合）**
    - **教育培训、电子游戏，娱乐服务中**
    - **智能计算机的人机界面**

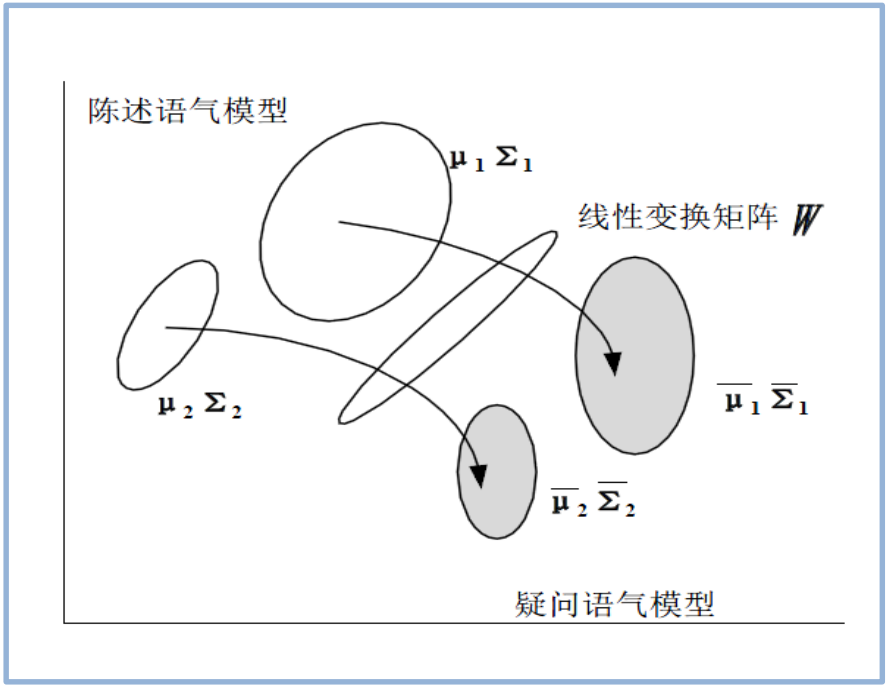
语音合成技术

语言合成技术之六 疑问句、感叹句、情感语音

- 采用极少语料量从陈述句训练的模型中自适应出疑问句感叹句口气



疑问



感叹

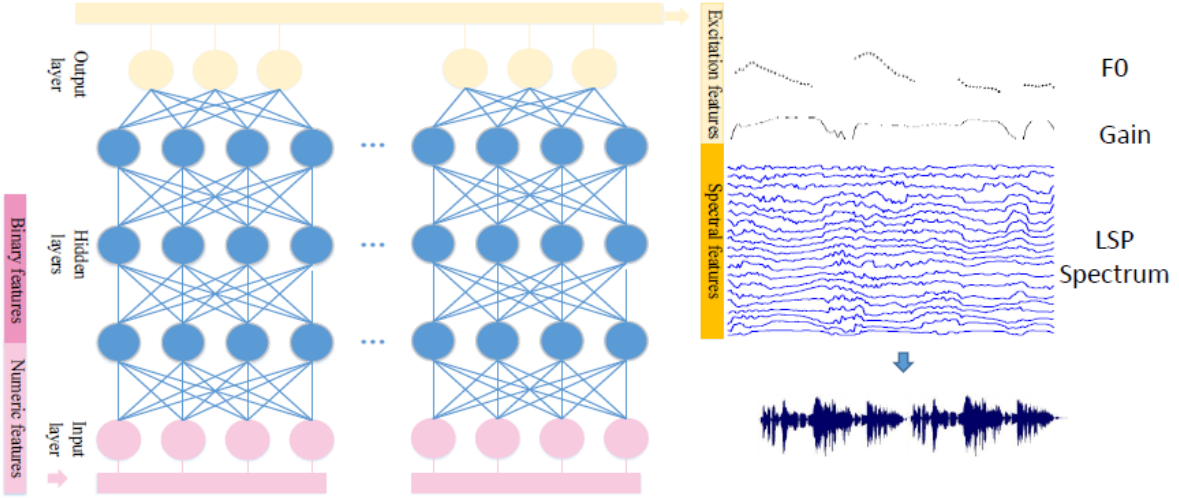
情感语音合成

	Sample 1	Sample 2	Sample 3
中性:			
悲伤:			
生气:			
高兴:			
害怕:			

语音合成技术

语言合成技术之七 深度学习语音合成

- 优点
  - 语音合成音质保真度高、韵律表现力高
  - 端到端系统很容易将系统扩展到多种语言
  - 易于实现
- 缺点
  - 需要较大规模训练语料库
  - 计算开销大



语音编码技术

- **语音编码：在保持可以接受的失真的情况下，采用尽可能少的比特数表示语音。**

  脉冲编码调制

  自适应预测编码

  自适应变换编码

  线性预测编码

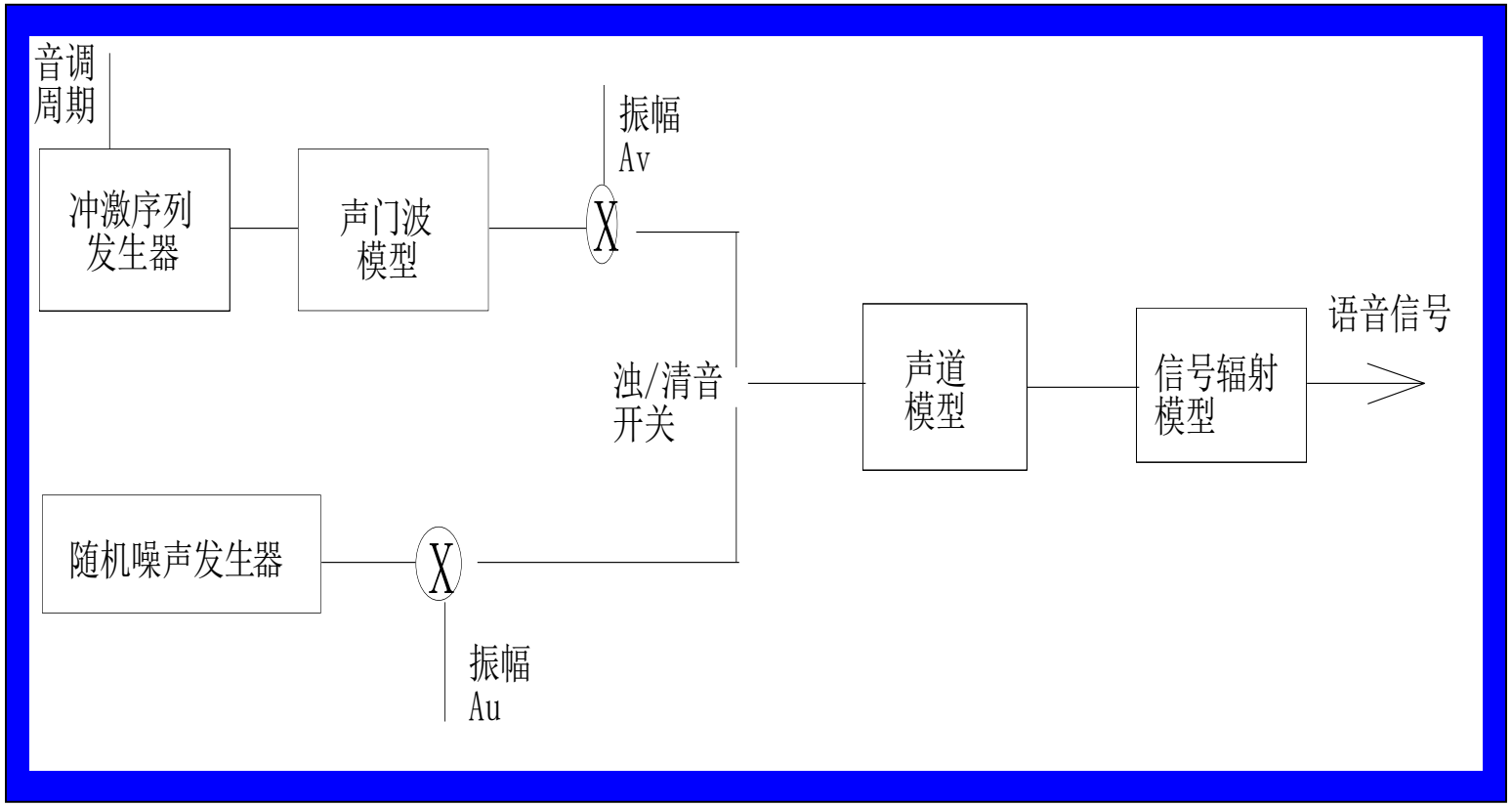
  线性预测声码器

  共振峰声码器

  相位声码器

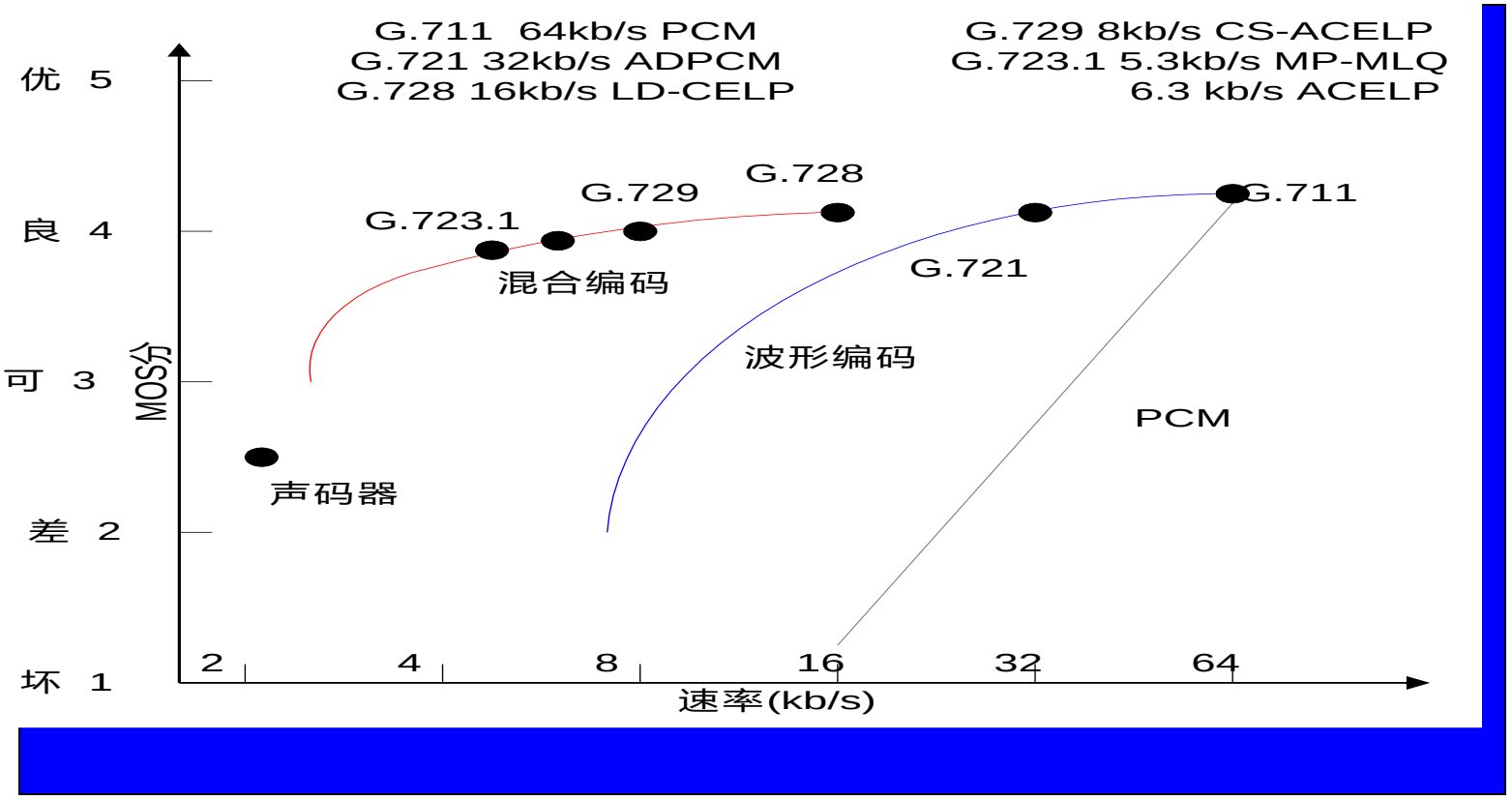
语音编码技术

产生语音信号的源-滤波器模型



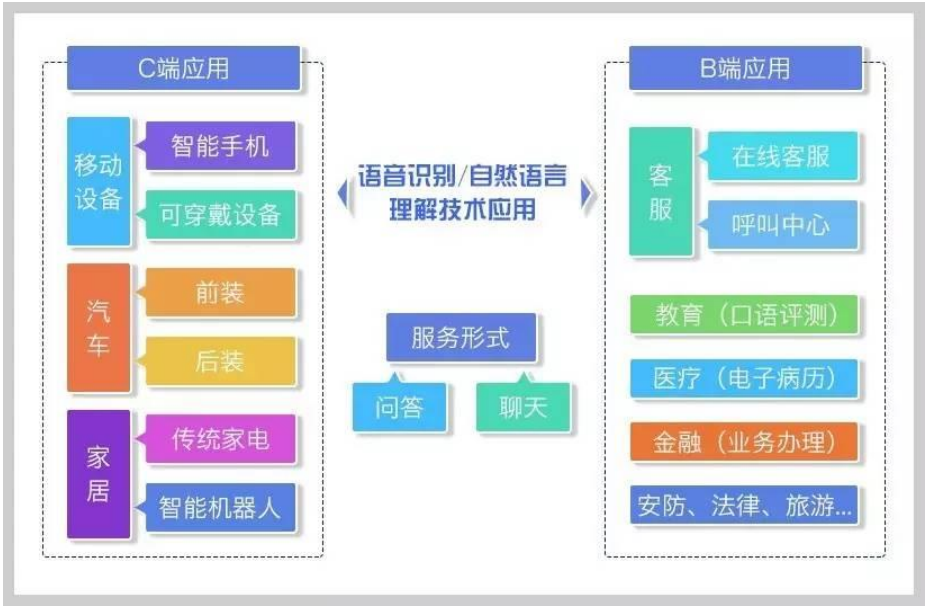
语音编码技术

语音压缩编码技术最新动态



语音技术应用

以问答和聊天为服务形式，智能语音语义在多个使用场景和行业领域都有广泛应用，我们可以简单从C端和B端两个方向分别来看。



C端应用方面，主要用于移动设备、汽车、家居三大场景，用来变革原有人机交互方式；B端则针对垂直行业需求，提升人工效率，比如帮助医生做电子病历录入，或代替部分人力工作，比如回答大部分简单重复的客服问题。由于两大领域解决的问题不同，因此遇到的挑战也各不相同。

语音技术典型应用（语音云+终端）

语音技术典型应用（终端）


TP memo
Info
Source
No traffic reports are available.
TP TMC 22:05

语音技术典型应用（人机交流）

本节课总结

- 语音基本概念
  - 语音产生、语音参数、计算机处理语音等基本概念
- 语音研究历史
  - 语音技术发展的不同时代的特点
- 语音技术概述
  - 重点分析语音识别（含声纹识别、语音增强等）、语音合成等技术的特点和最新发展情况

谢谢！