认知基础
Cognitive Foundation

第五章

听觉和言语
Audition and Speech

史忠植

中国科学院计算技术研究所

http://www.intsci.ac.cn/

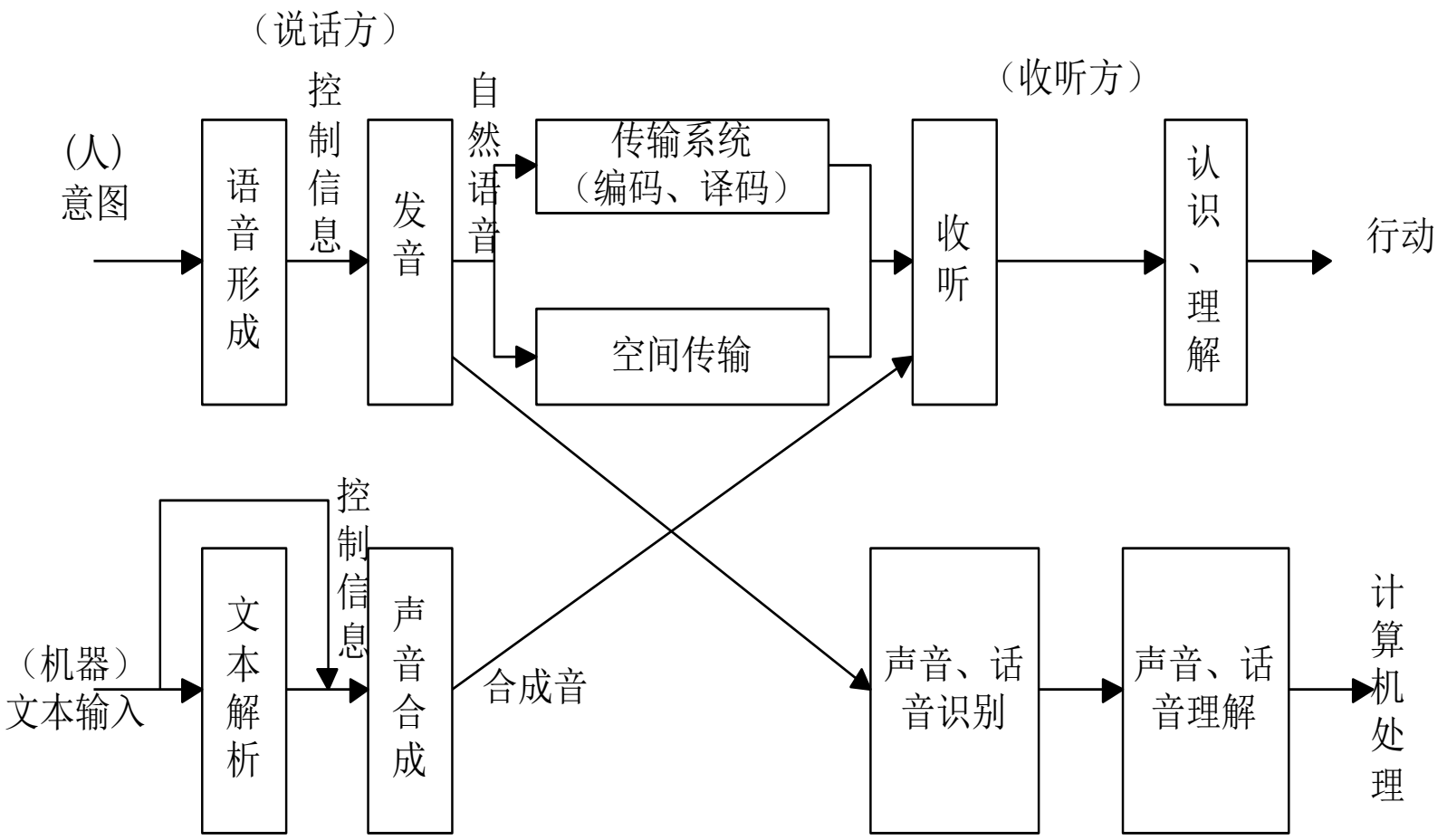
内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

听觉感知

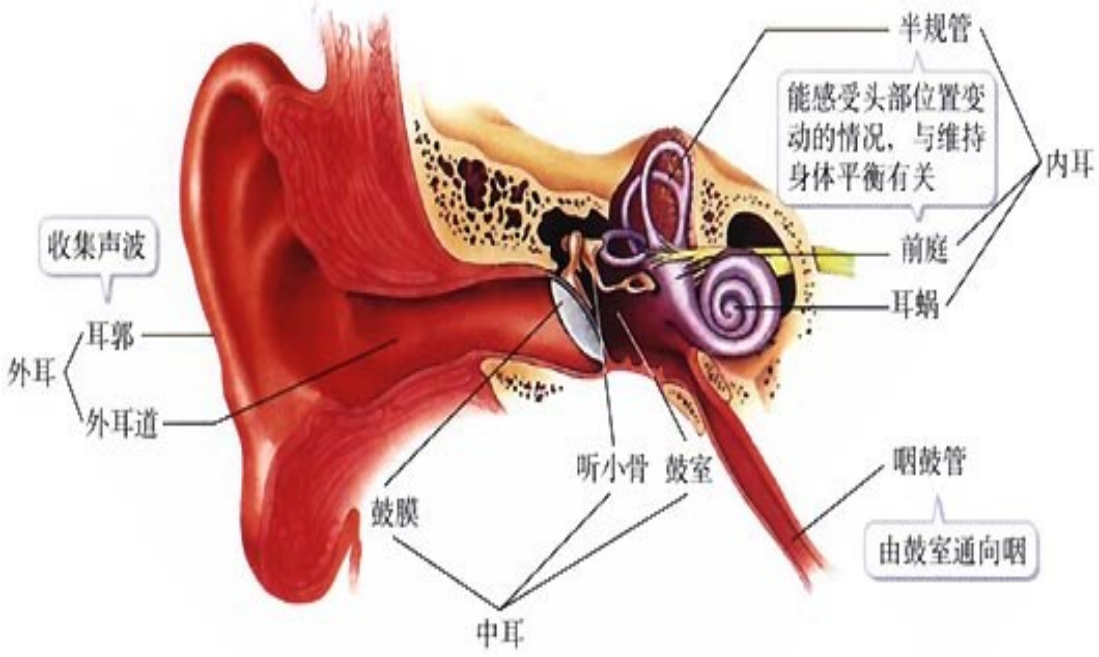
- 声波作用于听觉器官，使其感受细胞兴奋并引起听神经的冲动发放传入信息，经各级听觉中枢分析后会产生听觉。
- 听觉过程包括机械→电→化学→神经冲动→中枢信息加工等环节。从外耳的集声至内耳基底膜的运动是机械运动，毛细胞受刺激后引起电变化，化学介质的释放、神经冲动的产生等活动，冲动传至中枢后则是一连串复杂的信息处理过程。

交互信息流程图


（说话方）
控制信息
自然语音
（收听方）
(人)
意图
语音形成
发音
传输系统
（编码、译码）
空间传输
收听
认识、理解
行动
控制信息
（机器）
文本输入
文本解析
声音合成
合成音
声音、话音识别
声音、话音理解
计算机处理

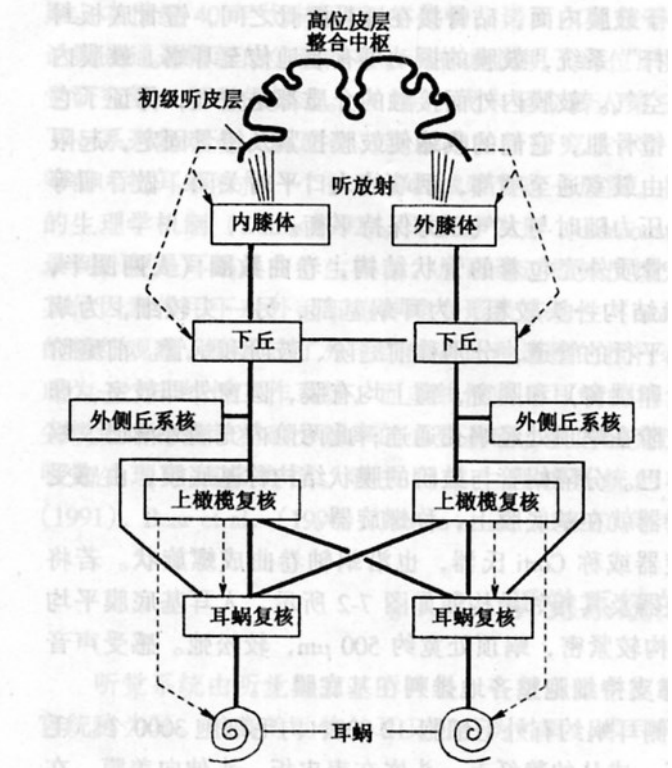
听觉形成的过程

- 口语一般由连续变化的声音模式以及少数停顿所组成。
- 声音→外耳→中耳→耳蜗听觉感受器→听神经→脑内听觉传导通路→大脑皮层听觉中枢一在此形成听觉。



听觉通路

从耳蜗到听觉皮质的听觉系统是所有感觉系统通路中最复杂的一种。听觉系统的每个水平上发生的信息过程和每一水平的活动都影响较高水平和较低水平的活动。在听觉通路中，从脑的一边到另一边有广泛的交叉。



听觉通路

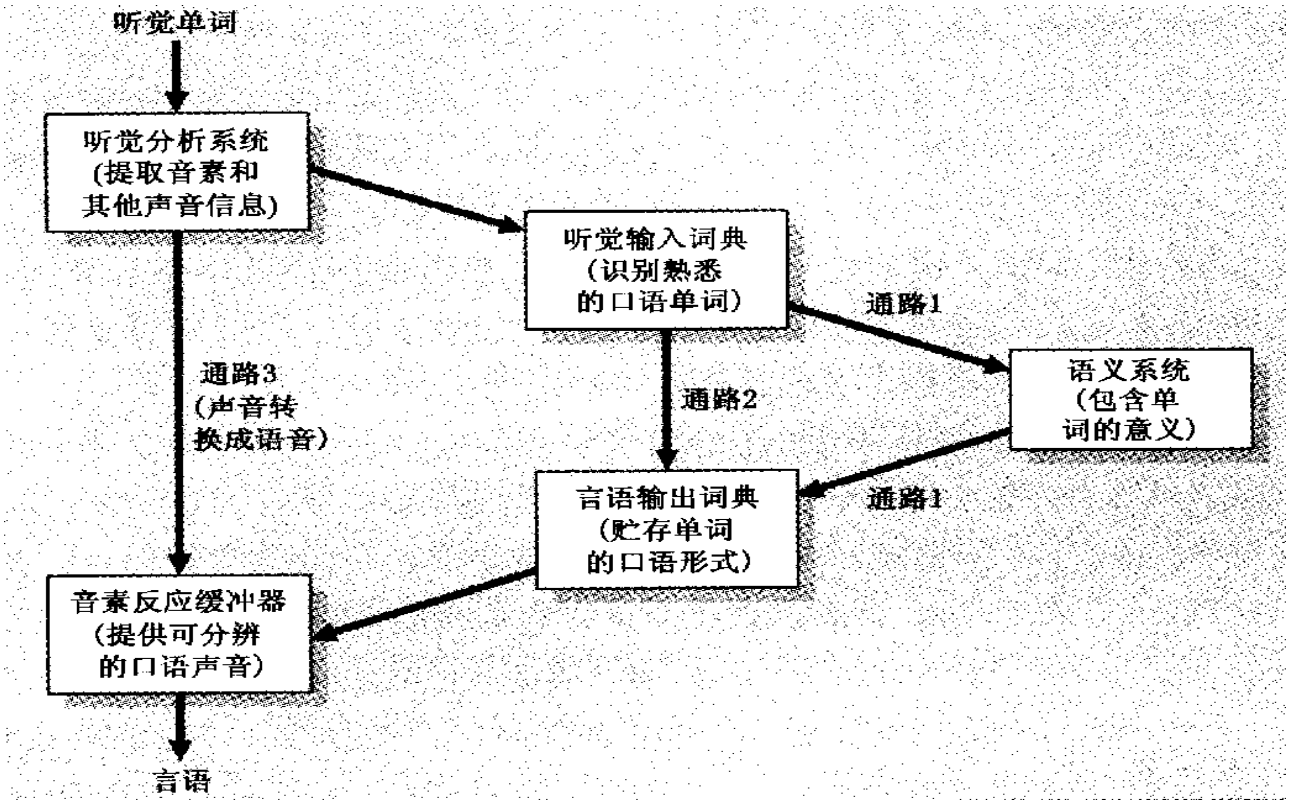
进入耳蜗神经核后，第八对脑神经听觉分支纤维终止于耳蜗核的背侧和腹侧。从两个耳蜗核分别发出纤维系统，从背侧耳蜗发出的纤维越过中线，然后经外侧丘系上升到皮质。外侧丘系最后终止于中脑的下丘，从腹侧耳蜗核发出的纤维，首先与同侧和对侧的上橄榄体复合体以突触联系。上橄榄体是听觉通路中的第一站，在这里发生两耳的相互作用。

听觉通路

- 上橄榄体复合体是听觉系统中令人感兴趣的中心，它由几个核组成，其中最大的是内侧上橄榄体和外侧上橄榄体。
- 从上橄榄复合体出发的纤维上升经过外侧丘系到达下丘。从下丘系将冲动传达到丘脑的内侧膝状体。连接这两个区域的纤维束，叫做下丘臂。从内侧膝状体，听觉反射的纤维将冲动传导颞上回（41区和42区），即听觉皮质区。

口语通路模型

- 1988年伊里斯（A.W. Ellis）和杨（A.W. Young）提出了一个口语单词加工的模型



口语通路模型

模型包括5个成分:

•听觉分析系统: 用于从声波中提取音素和其他声音信息。

•听觉输入词典: 包含听者知道的关于口语单词的信息，但不包含语义信息。这个词典的目的就是通过恰当地激活词汇单元来识别熟悉单词。

•语义系统：词义被贮存于语义系统之中。

•言语输出词典：用于提供单词的口语形式。

•音素反应缓冲器：负责提供可分辨的口语声音。

口语通路模型

听到一个单词至说出它之间存在三条不同的通路：

■ 通路1

这条通路利用听觉输入词典、语义系统和言语输出词典。它代表了无脑损伤人群正常识别和理解熟悉单词的认知通路。

■ 通路2

如果患者能够使用通路2，但通路1和3受到严重损伤，那么他们应该能够重复熟悉单词，但不能理解这些单词的意义。

口语通路模型

■ 通路3

如果一个患者只损伤通路3，那么他或她将展示在知觉和理解口语熟悉单词方面的完好的能力，但在知觉和重复不熟悉单词和非词时会出现障碍。这种情况临床上称之为听觉性语音失认。然而，他阅读非词时的能力完好。

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

听觉信息的中枢处理

中枢的信息处理过程甚为复杂，目前对它还缺乏较全面的理解。从耳蜗神经传入的冲动，在时间和空间上同所接受的声音特性不同而有不同的构型，这是输入信息编码的总形式，是当前听觉生理研究的核心问题。

- 频率分析机理
- 强度分析机理
- 声源定位和双耳听觉
- 对复杂声的分析

频率分析机理

自从一百多年前，亥姆霍兹(Helmholtz)提出了共振学说以来，不同的年代不同的作者提出过多种学说，解释耳蜗的频率分析机理。基本上可用两种观点进行概括：

- 部位学说：认为不同频率的声音兴奋基底膜不同部位的感受细胞，兴奋部位是频率分析的依据。
- 冲动频率学说：认为不同频率的声音使听神经兴奋后发放不同频率的冲动，冲动频率是声音频率分析的依据。

现在人们普遍认为它们二者不是互相排斥，而是可以互相补充的。

频率分析机理

- 行波论：行波论认为行波振幅最大点位置是对声音频率分析的依据，基底膜靠基部处接受高频声刺激，靠蜗顶处接受低频声刺激，当中按频率高低次序排列。因此基底膜就成为一个初级的频率分析器。
- 排放论：多根纤维随声波的周期而同步地轮流发放，则每一根纤维发放的频率不要很高，总体纤维上冲动排放却可跟上很高的频率。排放论认为听神经上冲动排放的频率与声音的频率是一致的，它是频率分析的依据。

频率分析机理

- 频率分析的中枢机理：

-从耳蜗核至听皮层的各级中枢的某些部位，神经细胞的排列都或多或少地有频率区域分布。如同视网膜的各部分在视皮层都有其对应的投射区似的，沿着耳蜗基膜的各部分在听觉皮层表面也有其系统性的特殊投射区。

-神经单元的放电与刺激声的同步锁相关系在内膝体及它以下各级听觉中枢中都被观察到，在耳蜗核和上橄榄核水平尤为明显。

-听觉中枢可能有积累传入信息并对它进行统计学处理的过程。

强度分析机理

感受细胞和神经单元的兴奋阈值有高有低，刺激强时被兴奋的感受细胞和神经单元便多，每一神经单元兴奋后发放神经冲动的数目也大。强度分析的依据：

- 被兴奋的单元是高阈值的还是低阈值的
- 兴奋单元的总数是多是少
- 发放的神经冲动是多是少

声源定位和双耳听觉

- 声源定位指听觉系统对声源方位的判断，它的基础是双耳听觉。由于从声源到两耳的距离不同及声音传波途中屏障条件的不同，从某一方位发出的声音到达两耳时便有时间差和强度差，它们的大小与声源的方位有关。双耳感受到的声音时间差和强度差便是声源定位的主要依据。对于高频声，强度差的作用较重要，对于低频声，时间差的作用较重要。
- 在双耳听觉的条件下，右耳对语言信号的感受似占较重要的地位，左耳则似对非语言信号的感受较重要，这可能和大脑两半球的分工有关。

对复杂声的分析

关于听觉系统如何辨别复杂声的问题，目前存在着两种截然不同的观点：

- 复杂声的感受以听觉系统对其简单组成成分的感受为基础，复杂声在听觉中枢引起的神经活动过程，是各组成成分引起的神经活动过程的总和；
- 听觉系统有分工检测各种复杂声音或声音某种特征的专门结构单元，称为探测器或特征探测器，它们只对特定的声音或特定的声音特征敏感，对其他声音或声音特征则无反应。

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

语音编码

语音编码方法归纳起来可以分成三大类：

- 波形编码：波形编码比较简单，编码前采样定理对模拟语音信号进行量化，然后进行幅度量化，再进行二进制编码。
- 信源编码：信源编码又称为声码器，是根据人声音的发声机理，在编码端对语音信号进行分析，分解成有声音和无声音两部分。
- 混合编码：混合编码是将波形编码和声码器的原理结合起来，数码率约在4kbit/s—16kbit/s，音质比较好。

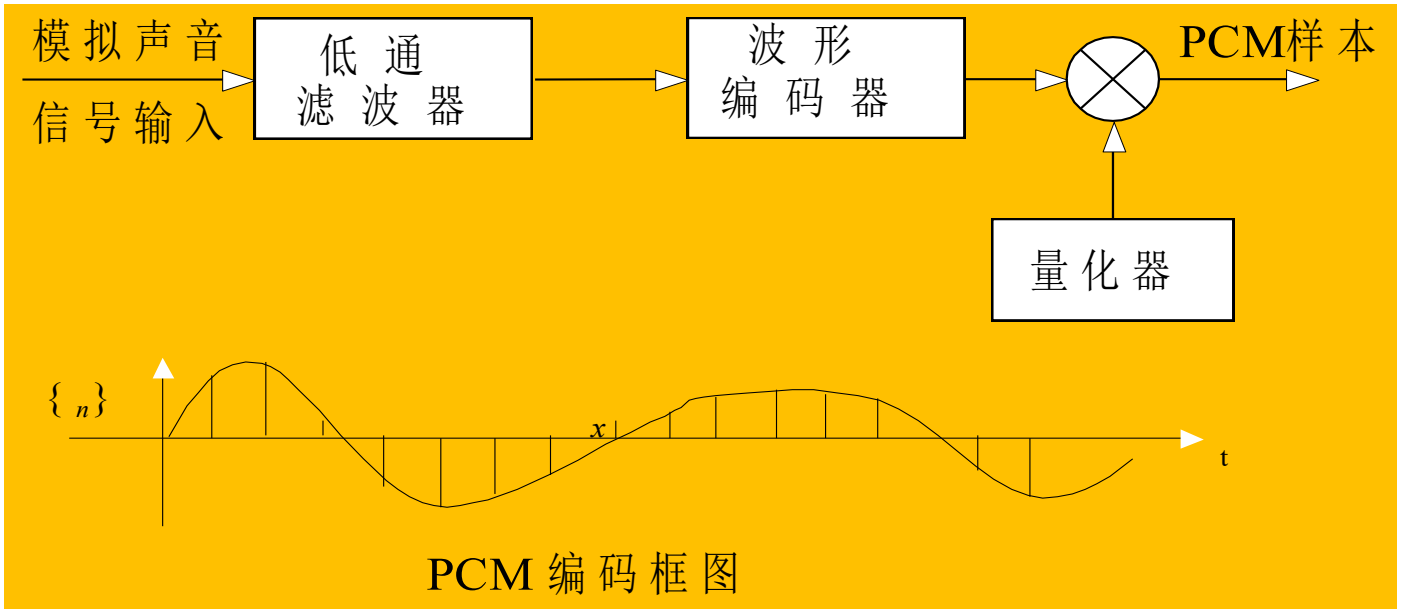
语音数字化的技术

波形编码的语音数字化最常用的技术：

- 脉冲编码调制（PCM）
- 差分PCM（DPCM）
- 增量调制（DM）

脉冲编码调制 PCM

在一定的时间间隔内，连续测量信号的幅度值，并对测量值编码。



PCM 编码框图

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

韵律认知

- 韵律是所有自然口语的共同特征，在言语交流中起着非常重要的作用，它通过对比组合音段信息，使说话者的意图得到更好的表达和理解。对人工合成语言而言，韵律控制模型的完善程度，决定了合成语言的自然度。
- 深入全面的理解自然语言的韵律特征无论对语音学研究，还是对提高语音合成的自然度和识别语音的准确性来说，都是至关重要的。语音流信息包括音段信息和韵律信息。音节等音段信息通过音色来表达，韵律信息则通过韵律特征来表达。

汉语的重音

韵律特征主要包含3个方面：重音、语调和韵律结构。

- 词重音和语句重音
- 语句重音的类型
- 语句重音的位置分布及等级差异

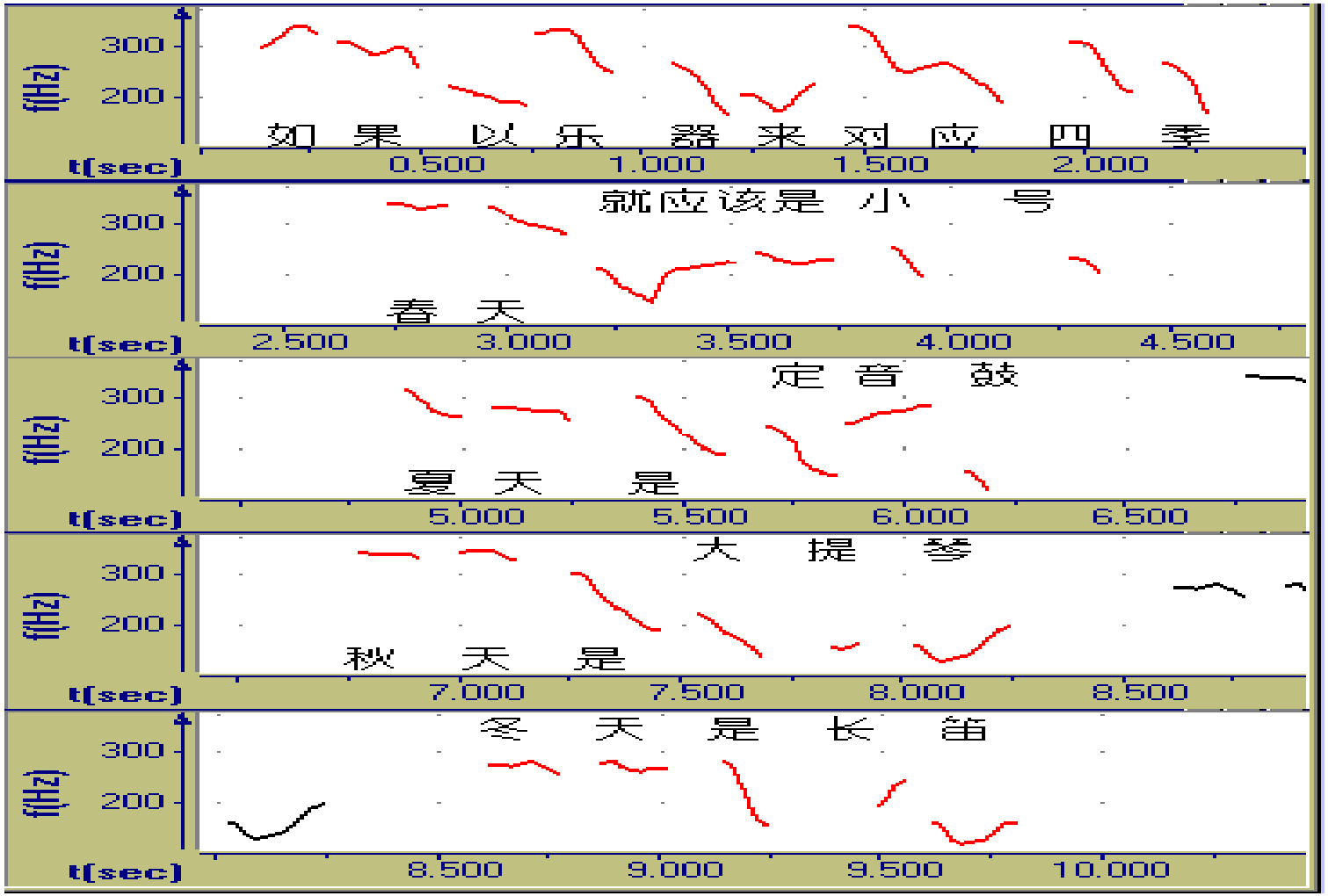
汉语韵律重音是复杂的：

“如果以乐器来对应四季，

春天就应该是小号，夏天是定音鼓，

秋天是大提琴，冬天是长笛。

汉语的重音


f(Hz)
300
200
t(sec)
0.500
1.000
1.500
2.000
如 果 以 乐 器 来 对 应 四 季
2.500
3.000
3.500
4.000
4.500
就应该是 小 号
春 天
5.000
5.500
6.000
6.500
定 音 鼓
夏 天 是
7.000
7.500
8.000
8.500
大 提 琴
秋 天 是
8.500
9.000
9.500
10.000
冬 天 是 长 笛

汉语语调

- 语调构造由语势重音配合而形成。它是一种语音形式，它通过信息聚焦来实施超语法的功能语义。节奏从语言的树型关系出发，按照表达的需要，利用有声形态有限的分解度来安排节奏重音，形成多层套叠的节奏单元。
- 语势重音和节奏重音分别调节声调音域的高音线和低音线。把语调构造各部分的音域特征综合在一起，可以区分出不同的语调类型来，因此语调是声调音域再调节的重要因素，而声调音域是功能语调和口气语调最重要的有声依据。

汉语语调

汉语语调的基本骨架可以分为调冠、调头、调核、调尾四部分。语调构造典型的有声表现如下

（1）调头和调核里有较强的语势重音，调冠里只有轻化音节，调尾里一般没有太强的语势重音。

（2）调核之后声调音域的高音线下移，形成明显的落差。

（3）调核后一音节明显轻化。调核为上声本调时，后一音节轻化并被明显抬高，高音线落差在其后出现。

韵律结构

韵律结构是一个层级结构，对它的成分有各种划分方法，一般公认有4个层级，从小到大依次是韵律词、韵律词组、韵律短语和语调短语：

- 韵律词：两音节或三音节组，在韵律词内部不能停顿
- 韵律词组：内部的韵律词之间没有停顿
- 韵律短语：韵律短语之间有比较明显的停顿
- 语调短语：在语调短语后一般有个比较长的停顿

韵律生成

- 迄今为止最全面的韵律产生模型是由莱弗特(W J M Levelt)等人提出来的韵律编码和加工模型。莱弗特认为口语句子的产生过程中，所有阶段的加工都是并行的、递增的。韵律编码包括词和在句子范畴的加工。
- 在一个句子的句法结构展开的同时，词汇的语音规划也产生了。词汇的通常分成两部分，lemma(包含语义和句法特征)的提取和lexeme (包含词形及音韵形式)的提取。后者由词形一韵律提取阶段执行，它用lemma作为输入来提取相应的词形和韵律结构。

韵律生成

- 在最后一个阶段，韵律产生器执行话语语音规划，产生句子的韵律和语调模式：
  - 产生韵律词、韵律短语和语调短语等韵律单元。词形一一韵律提取阶段的加工结果与连接成分组合，成为韵律词。通过扫描句子句法结构，再综合各种相关信息，然后把语法短语的扩展成分包含进来，组成一个韵律短语。而说话者语流某个点上的停顿，产生语调短语。
  - 在句子韵律结构和单个词的节律栅的基础上，韵律产生器最终构建出整个话语韵律结构的节律栅。

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

语音识别

自动语音识别（automatic speech recognition，ASR）是实现人机交互尤为关键的技术，让计算机能够“听懂”人类的语音，将语音转化为文本。自动语音识别技术经过几十年的发展已经取得了显著的成效。近年来，越来越多的语音识别智能软件和应用走人了大家的日常生活，苹果的Siri、微软的小娜（Cortana）、百度度秘（Duer）、科大讯飞的语音输入法和灵犀等都是其中的典型代表。随着识别技术及计算机性能的不断进步，语音识别技术在未来社会中必将拥有更为广阔的前景。

语音识别研究的历史

- 语音识别的研究起源于上世纪50年代，当时的主要研究者是贝尔实验室。早期的语音识别系统是简单的孤立词识别系统，例如1952年贝尔实验室实现了十个英文数字识别系统。从上世纪60年代开始，CMU的Reddy开始进行连续语音识别的开创性工作。
- 上世纪70年代，计算机性能的大幅度提升，以及模式识别基础研究的发展，例如码本生成算法（LBG）和线性预测编码（LPC）的出现，促进了语音识别的发展。
- 上世纪80年代是语音识别快速发展的时期，其中两个关键技术是隐马尔科夫模型（HMM）的理论和应用趋于完善以及NGram语言模型的应用。

语音识别研究的历史

- DARPA(Defense Advanced Research Projects Agency)是在1970年代由美国国防部远景研究计划局资助的一项10年计划，其旨在支持语言理解系统的研究开发工作。
- 到了1980年代，美国DARPA又资助了一项为期10年的DARPA战略计划，其中包括噪声下的语音识别和会话（口语）识别系统，识别任务设定为“（1000单词）连续语音数据库管理”。到了90年代，这一DARPA计划仍在持续进行中。其研究重点已转向识别装置中的自然语言处理部分，识别任务设定为“航空旅行信息检索”。

语音识别研究的历史

- 日本也在1981年的第五代计算机计划中提出了有关语音识别输入-输出自然语言的宏伟目标，虽然没能实现预期目标，但是有关语音识别技术的研究有了大幅度的加强和进展。
- 1987年起，日本又拟出新的国家项目---高级人机口语接口和自动电话翻译系统。

语音识别研究的历史

- 语音识别开始从孤立词识别系统向大词汇量连续语音识别系统发展。其核心框架就是用隐马尔科模型对语音的时序进行建模，而用高斯混合模型（GMM）对语音的观察概率进行建模。基于GMM-HMM的语音识别框架在此后很长一段时间内一直是语音识别系统的主导框架。
- 关键突破起始于2006年。这一年辛顿（Hinton）提出深度置信网络（DBN），促使了深度神经网络（Deep Neural Network，DNN）研究的复苏，掀起了深度学习的热潮。2009年，辛顿以及他的学生默罕默德（D. Mohamed）将深度神经网络应用于语音的声学建模，在小词汇量连续语音识别数据库TIMIT上获得成功。

我国语音识别研究的历史

- 我国的语音识别研究起始于1958年，由中国科学院声学所利用电子管电路识别10个元音。直至1973年才由中国科学院声学所开始计算机语音识别。由于当时条件的限制，我国的语音识别研究工作一直处于缓慢发展的阶段。
- 进入1980年代以后，随着计算机应用技术在我国逐渐普及和应用以及数字信号技术的进一步发展，国内许多单位具备了研究语音技术的基本条件。

我国语音识别研究的历史

- 1986年3月我国高科技发展计划(863计划)启动，语音识别作为智能计算机系统研究的一个重要组成部分而被专门列为研究课题。在863计划的支持下，我国开始了有组织的语音识别技术的研究，并决定了每隔两年召开一次语音识别的专题会议。从此我国的语音识别技术进入了一个前所未有的发展阶段。
- 在北京有中科院声学所、自动化所、清华大学、北京交通大学等科研机构和高等院校。另外，还有北京大学、哈尔滨工业大学、中国科技大学、四川大学等也纷纷行动起来。

我国语音识别研究的历史

- 现在，国内有不少语音识别系统已研制成功。这些系统的性能各具特色。

– 在孤立字大词汇量语音识别方面，最具代表性的要数1992年清华大学电子工程系与中国电子器件公司合作研制成功的THED-919特定人语音识别与理解实时系统。

– 在连续语音识别方面，1991年12月四川大学计算机中心在微机上实现了一个主题受限的特定人连续英语---汉语语音翻译演示系统。

– 在非特定人语音识别方面，有清华大学计算机科学与技术系在1987年研制的声控电话查号系统并投入实际使用。

语音信号

- 语音：人们讲话时发出的话语叫语音。是一种人们进行信息交流的声音，是组成语言的声音/带有语言信息的声音。
- 语音(Speech)=声音(Acoustic)+语言(Language)

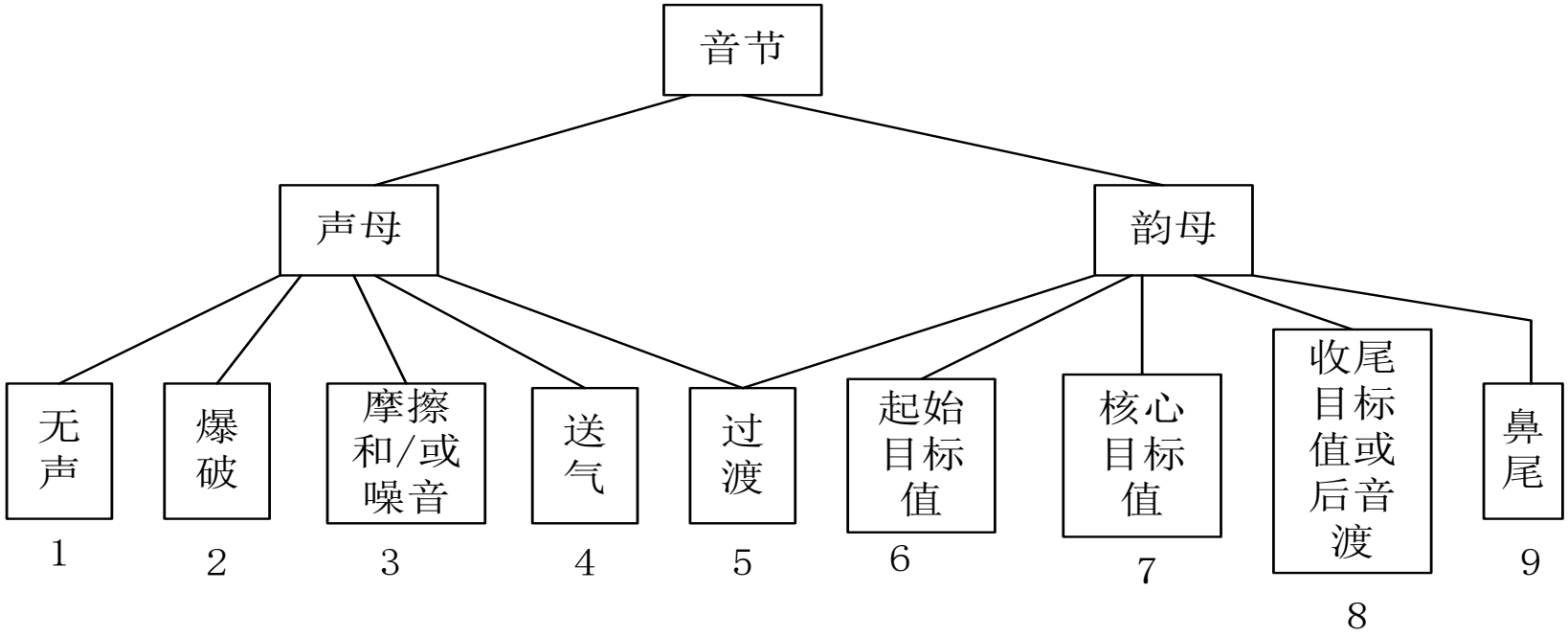
  语音是由一连串的音素组成语言的声音。

音节和音素

- 语音信号的最基本组成单位是音素，音素可以分为浊音和清音，在短时分析的基础上可以判断一段语音属于哪一类。
- 浊音短时谱的特点：
  - 具有明显的周期性起伏结构
  - 具有明显的凸起点，称为“共振峰”（formant）
- 清音短时谱的特点：随机噪声

音　节

- 音节是发声的最小单位，一个音节由元音和辅音构成，“辅音-元音”



音节和音素

- 语音信号的最基本组成单位是音素，音素可以分为浊音和清音，在短时分析的基础上可以判断一段语音属于哪一类。
- 浊音短时谱的特点：
  - 具有明显的周期性起伏结构
  - 具有明显的凸起点，称为“共振峰”（formant）
- 清音短时谱的特点：随机噪声

声　道

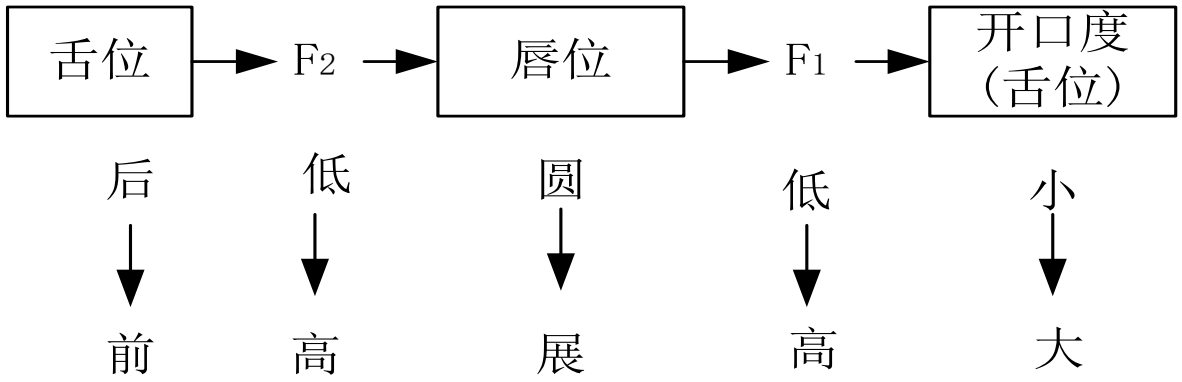
- 不同的韵母是由于声道形状的不同造成的，声道可以用一段变截面积的声管来表示。
- 声道形状主要取决于三个方面：
  - 舌在口腔中的前后位置不同，造成收紧点(面积最小点)的位置不同
  - 舌位的高低，舌位越高嘴张的越大，也称开口度大
  - 唇的圆展程度

元音（汉语中称为韵母）

- 单韵母5个，[a]，[i]，[u]，[ü]， [e]，[o]
- 复韵母14个，[ai]，[ei]，[au]，[ou]，[ia]，[ie]，[ua]，[uo]，[üe]，[iao]，[uai]，[uei]
- 鼻韵母16 个，[an]，[ian]，[uan]， [üan]，[en]，[in]，[uen]，[ün]，[ang]，[iang]，[uang]，[eng]，[ing]，[ueng]，[ong]，[iong]

舌位与口型

- 舌位的前后主要影响第二共振峰，舌位靠前，收紧点靠前，第二共振峰越高。
- 舌位上下即开口度主要影响第一共振峰，开口度越小，第一共振峰越低。
- 唇的圆展程度对第一共振峰和第二共振峰都有影响



单韵母发音及频谱特点

韵母	典型字的韵母	收紧点	开口度	F1	F2
[a]	巴、大	后	大	850	1300
[i]	一、希	前	小	300	2300
[u]	乌、路	后	小	350	650
[ü]	玉、居	前	小	300	2000
[e]	特、哥	中	中	520	1200
[o]	迫、魔	中	中	570	840

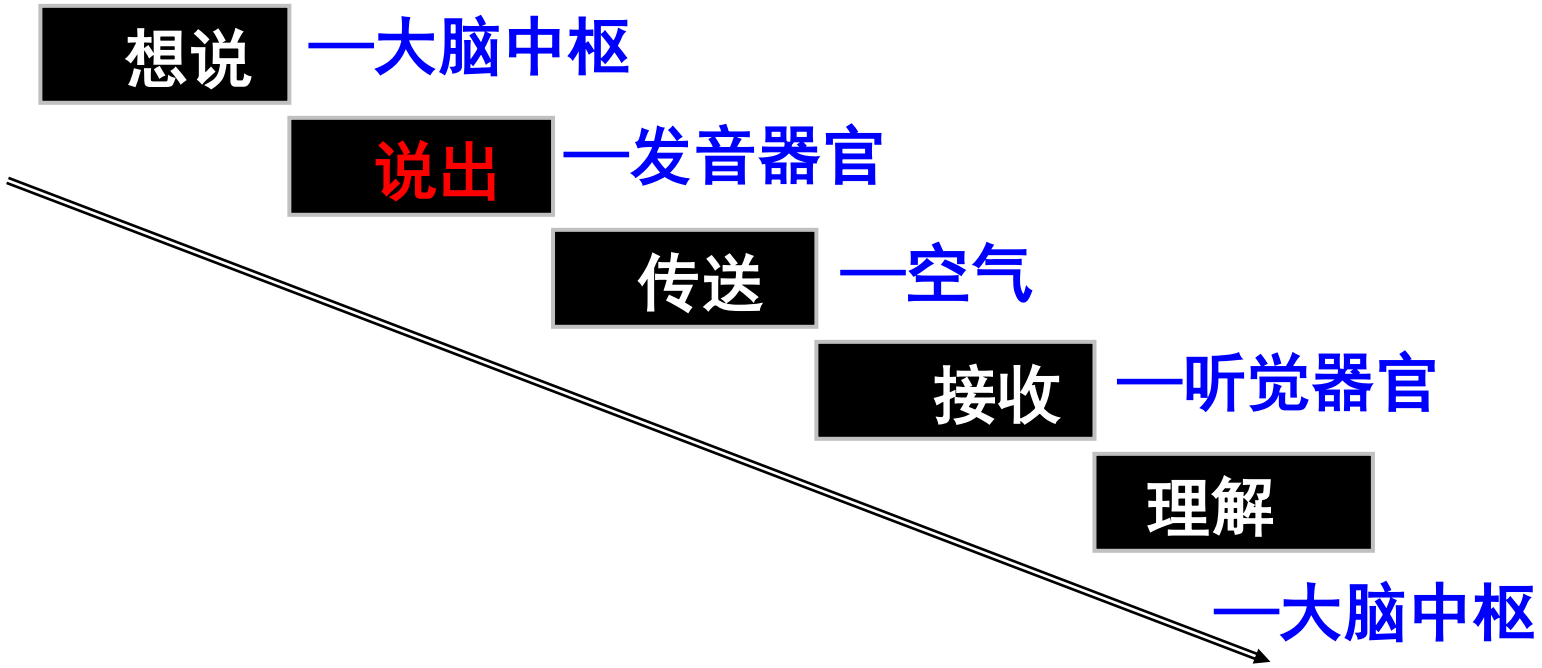
声　母

汉语普通话声母的划分：

- 不送气塞音 [b]，[d]，[g]
- 送气塞音[p]，[t]，[k]
- 清擦音[s]，[sh]，[x]，[f]，[h]
- 不送气塞擦音[z]，[zh]，[j]
- 送气塞擦音[c]，[ch]，[q]
- 鼻音[m]，[n]
- 边音[l]
- 卷舌音[r]

说话过程

- 人类的说话过程分五个阶段



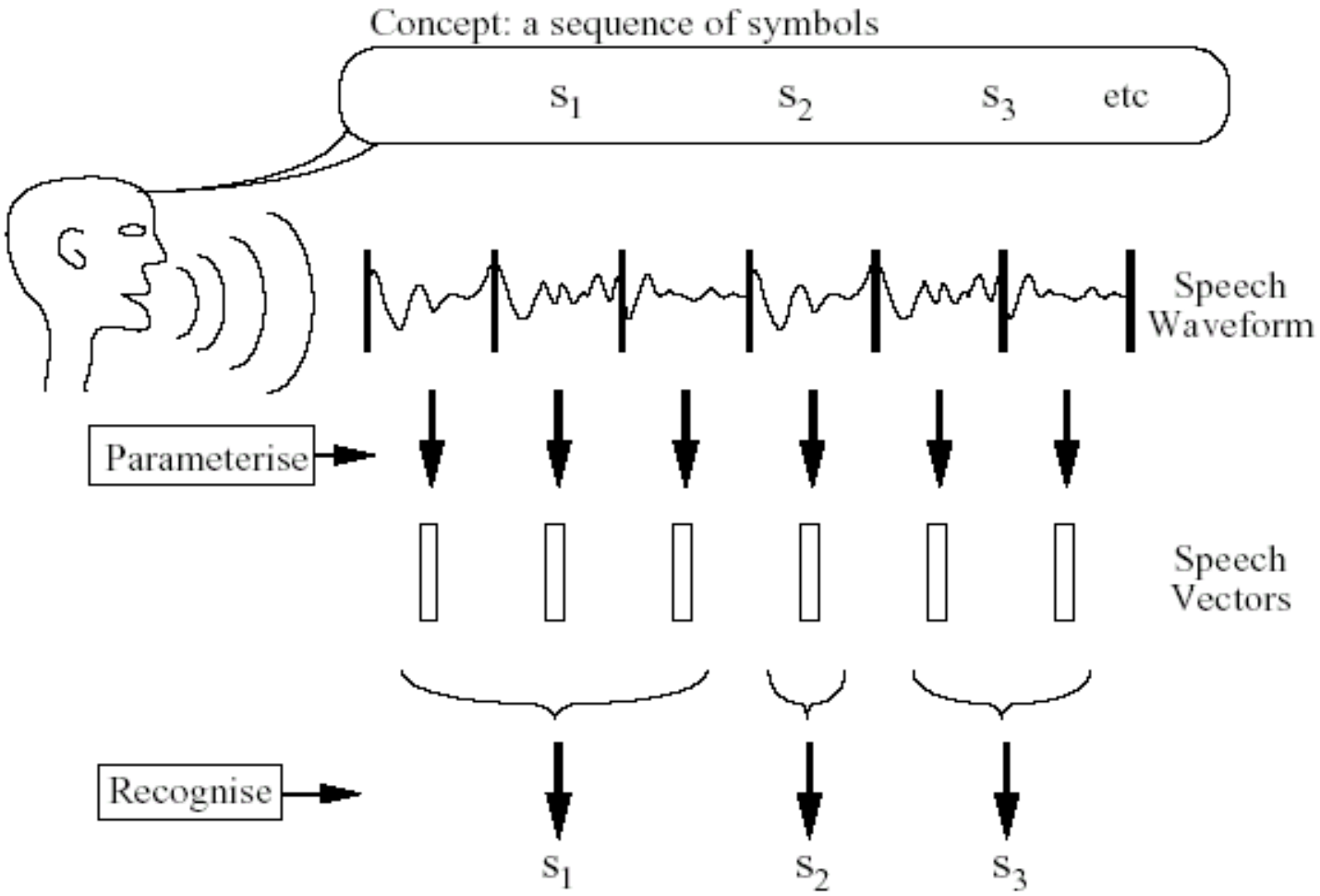
语音识别基本术语

- 特定人和非特定人（话者相关或话者无关）
- 词汇量（大，小）
- 孤立词，连接词，关键词和连续语音
- 自然发音和朗读发音
- 口音（方言）
- 背景噪音（环境噪音）
- 信道差异（固定电话，麦克，手机等）
- 声学模型（HMM，mono-phone，bi-phone，tri-phone）
- 声学特征（MFCC）
- 解码（Viterbi）

语音识别指标

- 识别指标：
  - -SER（Sentence Error Rate，句子错误率）
  - -WER（Word Error Rate，词错误率）
  - -CER（Character Error Rate，字错误率）
  - -PER（Phone Error Rate，音节错误率）
- 采样率，8kHz（电话或手机），16kHz（麦克风）
- 时域，频域
- 端点检测，静音检测或有效音检测（VAD）

语音识别过程

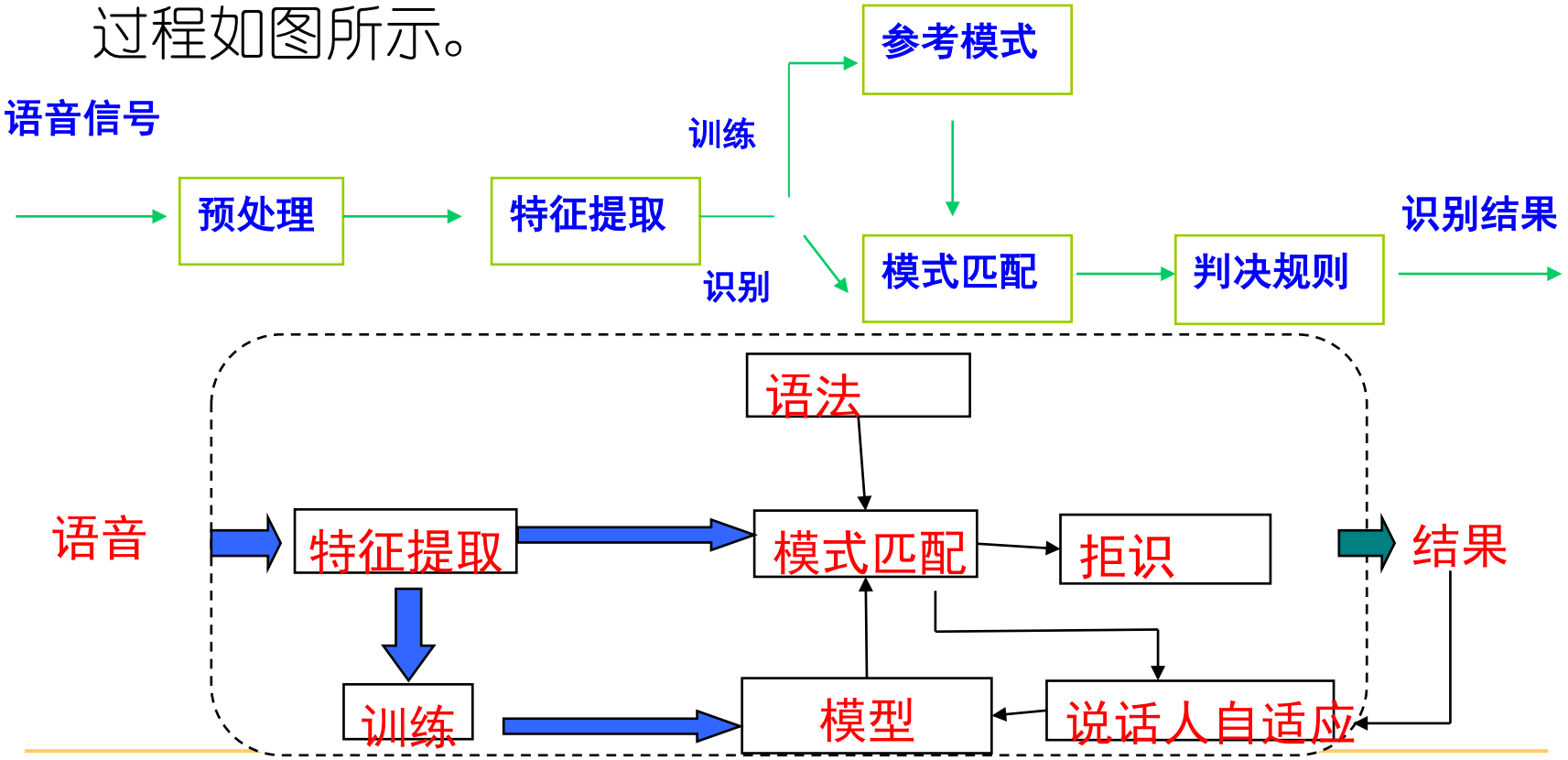

Concept: a sequence of symbols
s_1 s_2 s_3 etc
Speech Waveform
Parameterise
Speech Vectors
Recognise
s_1
s_2
s_3

语音识别的基本原理

- 训练(Training)：预先分析出语音特征参数，制作语音模板(Template)并存放在语音参数库中。
- 识别(Recognition)：待识语音经过与训练时相同的分析，得到语音参数，将它与库中的参考模板一一比较，并采用判决的方法找出最接近语音特征的模板，得出识别结果。
- 失真测度(Distortion Measures)：在进行比较时要有个标准，这就是计量语音特征参数矢量之间的“失真测度”。
- 主要识别框架：基于模式匹配的动态时间规整法(DTW:Dynamic Time Warping)、基于统计模型的隐马尔柯夫模型法(HMM:Hidden Markov Model)、深度学习。

语音识别原理框图

- 不同的语音识别系统，虽然具体实现细节有所不同，但所采用的基本技术相似，一个典型语音识别系统的实现过程如图所示。

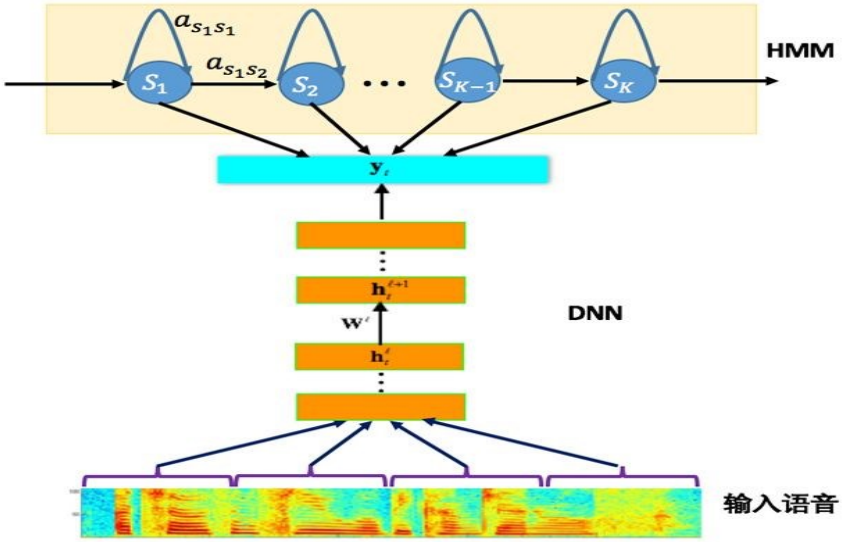


基于深度学习的语音识别

- 2011年，微软研究院俞栋、邓力等发表深度神经网络在语音识别上的应用文章，在大词汇量连续语音识别任务上获得突破。从此基于GMM-HMM的语音识别框架被打破，大量研究人员开始转向基于DNN-HMM的语音识别系统的研究。上世纪80年代是语音识别快速发展的时期，其中两个关键技术是隐马尔科夫模型（HMM）的理论和应用趋于完善以及NGram语言模型的应用。

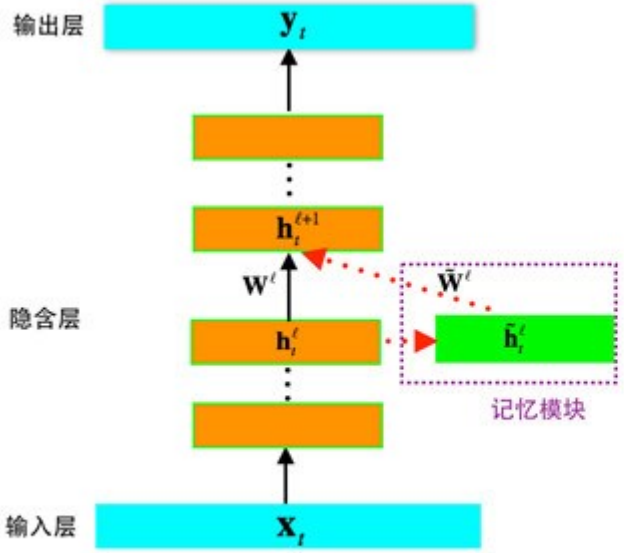
基于深度学习的语音识别系统

- 使用DNN估计HMM的状态的后验概率分布不需要对语音数据分布进行假设
- DNN的输入特征可以是多种特征的融合，包括离散或者连续的；
- DNN可以利用相邻的语音帧所包含的结构信息。

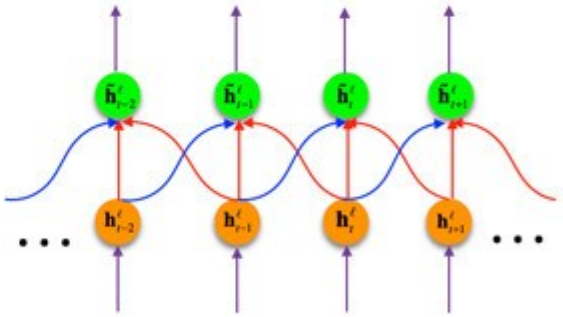


科大讯飞DFCNN语音识别框架

- 科大讯飞科大讯飞研发了一种名为前馈型序列记忆网络FSMN (Feed-forward Sequential Memory Network) 的新框架。FSMN可和CTC（Connectionist temporal classification）准则结合，实现语音识别中的“端到端”建模。



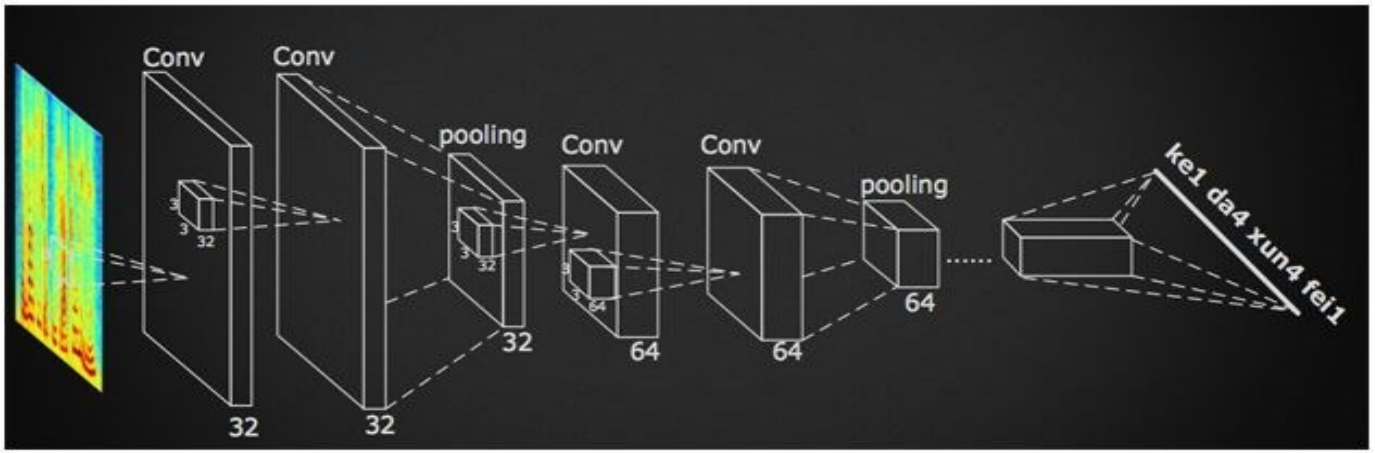
(a) FSMN 结构示意图



(b) 第ℓ个隐层记忆模块的时序展开示意图（左右各看1帧）

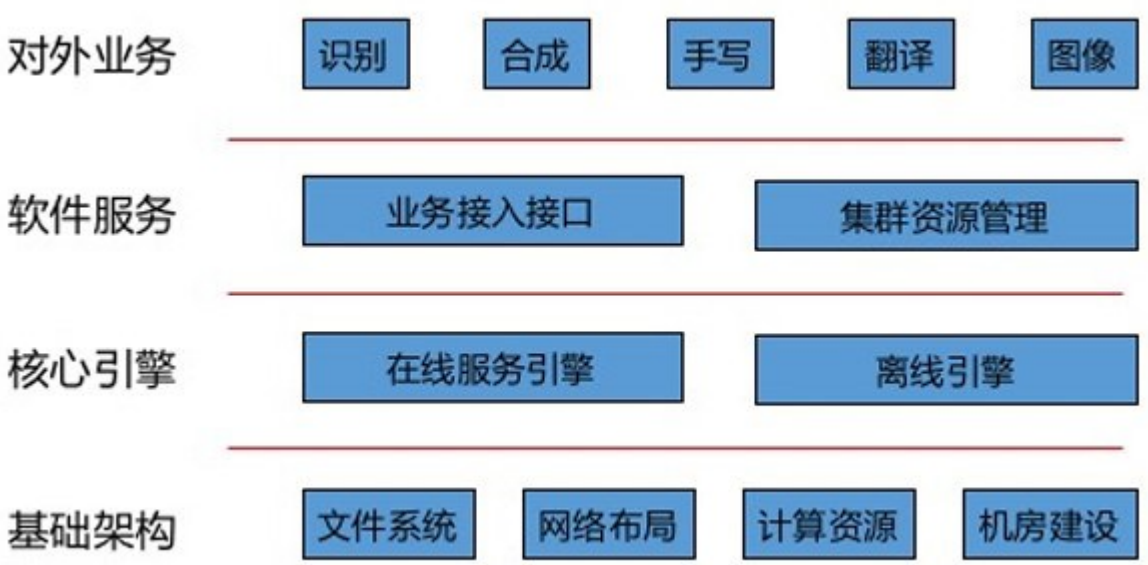
科大讯飞DFCNN语音识别框架

- 科大讯飞深度全序列卷积神经网络（Deep Fully Convolutional Neural Network，DFCNN）的语音识别框架，使用大量的卷积层直接对整句语音信号进行建模，更好地表达了语音的长时相关性。



科大讯飞深度学习平台

- 科大讯飞分析算法的计算特点，搭建了一套快速的深度学习计算平台——深度学习平台。



- 整个平台分为四个组成部分：底层基础架构、核心计算引擎、软件服务、对外业务。

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

语音合成

- 语音合成（Speech Synthesis） 是讨论如何使机器说出人的语言，以满足人类的各种需要的问题。
- PCM波形合成法即分析-存储-合成”。在数字语音合成中，为了便于存储，必须先进行分析或变换，因而在取出合成前还必须进行相应的反变换，最简单的变换是模数变换和数模变换。这种方法合成语音，其词汇量不能很大，所需的存储空间太大。如要让机器讲1秒钟的语音，就需要64kbit以上的存储容量。
- 参数合成法。为了节约存储量，必须先对语音信号进行各种分析，得到诸如线性预测系数、线谱对参数或共振峰参数等有限个参数，以压缩存储容量。

语音合成

- 在目前的技术水平下，要想合成任意一语种的无限词汇量的语音，仅采用上述的“分析-存储-合成”法是不可能的。
- 甚至对于以音节为基础，且字汇量较少的汉语，若以音节字为合成基元，也有1300个音节字，即使使用参数存储也将是很困难的。
- 因此国际上都在努力开发另一类无限（全）词（字）汇量的语音合成方法，这第二类法就是所谓“按语言学规则的从文本至语言”的语言合成法（Test- to - Speech Synthesis by Rule）。

语音合成

- 我国的汉语，在无限字（词）汇量的语音合成，具有得天独厚的优越性。
- 汉语的句子是由词组成的，而词又是由音节字组成的。
- 虽然存在一音多字的问题，但是对于机器讲话、人听话的语音合成情况来说，这个同音字问题是不必考虑的。因为人在听话时会自然的理解这些同音字，也就是说，汉语合成时只是要求机器讲出音节字（拼音字）就可以了。
- 汉语的全部音节字只有1300个左右，即使不用更小的声母、韵母作为基元就用音节字作为基元，其语音库也不算太大。

语音合成

- 语音合成技术可以分为四类：
  - -波形编码合成方法 (Waveform Coding Synthesis)
  - -参数式分析合成方法 (Parametric Analysis Synthesis)
  - -规则合成方法 (Synthesis by Rule)
  - -文－语转换 (Text to Speech Conversion System)
- 无论波形合成法或是参数合成法，其原理都等同于语音通信的语音编码或声码器中的接收端的工作过程，只是现在没有从信道送来的参数与编码的序列，而代之以从分析或变换得到的存储在语音库中的参数或码序列。

波形编码方法

这种方式以语句、短语、词或音节为合成单元，这些单元分别被录音后直接进行数字编码，经过适当的数据压缩，组成一个合成语音库。

- 重放时，根据待输出的信息，在语音库中取出相应单元的波形数据，串接或编辑在一起，经过解码还原出语音，这种合成方法也叫录音编辑合成。
- 合成单元越大， 合成的自然度越好，系统结构简单，价格低廉，合成语音的数码率较大，存储量也大，因而合成的词汇有限。
- 在自动报时、报号、报站或报警等装置中，多采用这种技术，现有多种合成芯片可供选用。

参数式分析合成

- 这种合成方法多以音节、半音节或音素为合成单元。
- 首先按照语音理论，对所有合成单元的语音进行分析，一帧一帧提取有关语音参数，这些参数经编码后组成一个合成语音库。
- 输出时，根据待合成的语音的信息，从语音库中提取出相应的合成参数， 经编辑和连接顺序送入语音合成器中，在合成器中合成参数的控制下，一帧一帧的重新还原语音波形。
- 主要的合成参数有：控制音强的幅度、控制音高的基频和控制音色的共振峰参数。
- 这种方式的速码率比波形编辑方式小的多，但是系统结构也复杂些，合成音质也差些。目前已有专用的芯片和界线板

规则合成

- 这种合成方法以通过语音学规则来产生任何语音为目的，规则合成系统存储的是较小的语音单位的声学参数以及由音素组成音节，再由音节组成词或句子的各种规则。当**输入字母符号**时，合成系统利用规则自动将它们转换为连续的语音声波。
- 由于语音中存在协同发声效应，单独存在的元音和辅音与连续发音中的元音和辅音不同，所以合成规则是在分析每一语音单元出现在不同环境中的协同发音后，归纳其规律而制定的如共振峰的频率规则、时长规则、声调和语调规则等。
- 与分析合成方法相比，规则合成方法的语音库的存储量更小，这是以牺牲音质为代价的，这种方式涉及到许多语音学和语音学模型，系统结构复杂。目前合成规则还不完善，合成音质一般较差。

文-语合成

- 文-语合成的指导思想是：挖掘出人在讲话时，是按照什么规则来组织语音单元的，并将这些规则的知识赋予机器，因而机器在合成语音时，只要**输入合成基元**，机器就应该会按照所赋给的规则来合成出与人讲话是相同的语音来。
- 应该指出，所使用的文本的合成基元越小，合成规则就越多越复杂，当然所用的存储量也就越小。因此在选择文本的合成基元时应该折衷考虑。
- 目前英语中多用音素、双音素为文本的合成基元，因为对于西方语言，用词作为基元的按规则合成几乎是不可能的。而汉语可以用声母和韵母，甚至直接用音节字作为文本基元，以减少规则的知识。这时就不必靠与音素有关的规则，而只需用到音节字之间的有关规则。

文—语转换系统

- 这是一种以文字串为输入的规则合成系统，其输入的文字串是通常的文本字串，系统中的文本分析器首先根据发音字典，将输入的文字串分解为带有属性标记的词及其读音符号，再根据语义规则和语音规则，为每个词、每个音节确定重音等级和语句结构及语调、以及各种停顿等，这样，文字串就变换为代码串，规则合成系统就可以据此合成抑扬顿挫和不同语气的语句。
- 文－语转换系统除了依赖各种规则（包括语义规则，词规则，语音学规则）外，还必须对文字内容有正确的理解，也就是自然语言理解问题，所以真正的文－语转换系统实际上是一个人工智能系统。迄今为止，还没有开发出一套相当满意的文－语转换系统。

三种语音合成方式特征比较

		波形合成法	参数合成法	规则合成法
基本信息		波形	特征参数	语言的符号组合
语音质量	可懂度	高	高	中
	自然度	高	中	低
词汇量		小(500字以下)	大(数千字)	无限
合成方式		PCM、ADPCM、APC	LPC、线谱对LSP、共振峰	LPC、线谱对LSP、共振峰
数码率		9.6～64 kbit/s	2.4～9.6 kbit/s	50～75 bit/s
1 Mbit可合成的语音长度		15～100 s	100s～7分钟	无限
合成单元		音节，词组，句子	音节，词组，句子	音素、音节
装置		简单	比较复杂	复杂
硬件主体		存储器	存储器和微处理器	微处理器

共振峰式语音合成器

- 音色各异的语音具有不同的共振峰模式，因此以每个共振峰频率及其带宽为参数，可以构成一个共振峰滤波器，用若干个这种滤波器的组合来模拟声道的传输频响，对激励源发出的信号进行调制，再经过辐射即可得到合成语音。
- 早期的共振峰滤波器是用模拟电路来实现的，现在都用数字滤波器来实现。

共振峰式语音合成器

- 共振峰语音合成原理如下：

①浊音和清音分别采用不同的激励源。

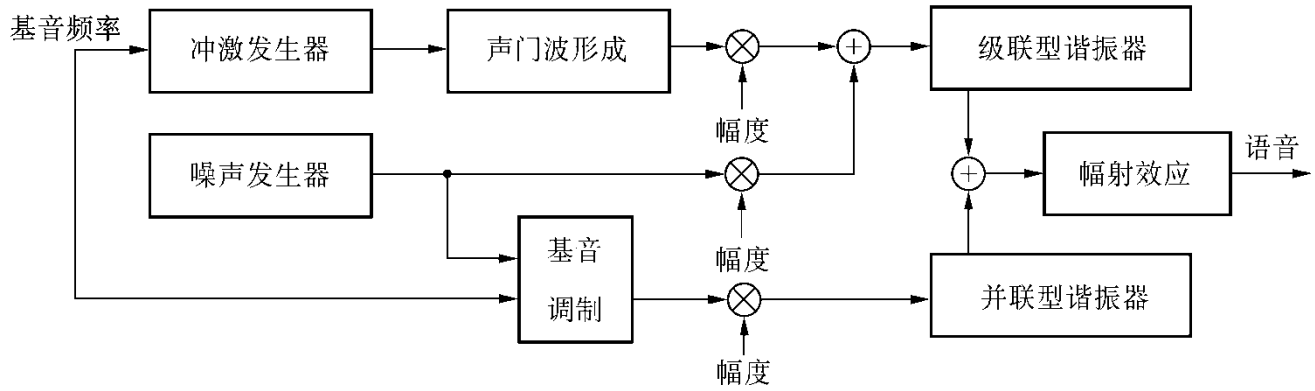
②发不同语音对应不同声道路径和滤波器。

③发不同语音对应不同幅值控制和频率控制。

④共振峰和基频是语音信号的2个主要特征。

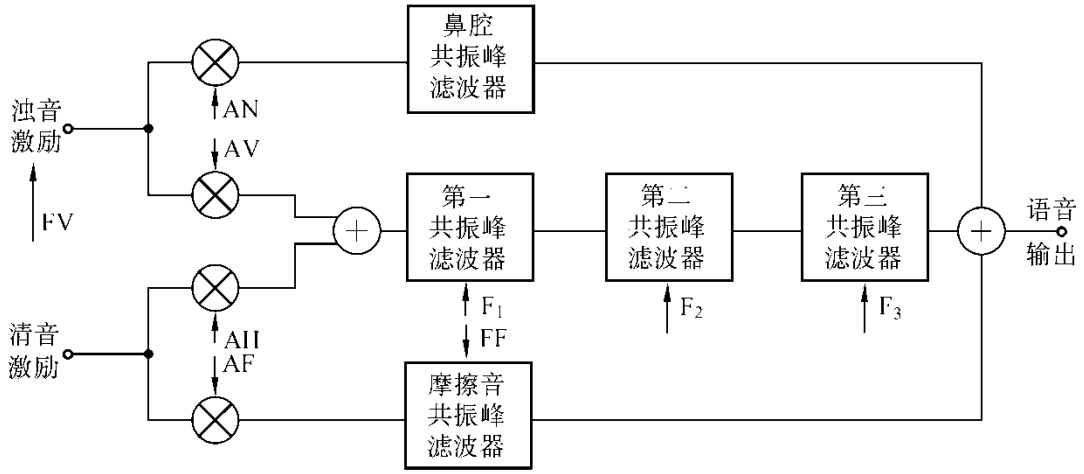
- Vortax公司推出的Computalker是一种典型的语音合成产品——最早进入计算机业余爱好者市场。采用的便是共振峰语音合成原理。

共振峰式语音合成器



- 激励源对合成语音的自然度有明显的影响，激励源有三种类型：
  - 合成浊音语音时用周期冲激序列，
  - 合成清音语音时用伪随机噪声，
  - 合成浊擦音时用周期冲激调制的噪声

共振峰式合成器实例



Computalker共振峰语音合成原理框图

- 中间的信号传输通道对应于口腔的发音，这是主要声道路径。元音和部分辅音通过此路径发音。
- 口腔语音不用鼻腔，而鼻音用口腔和鼻腔发音。因此发鼻音时要附加一并联于口腔的鼻腔，用一个鼻腔共振峰滤波器来模拟。
- 部分辅音如摩擦音的发音虽然也用口腔，但其共振峰不同，因此发这部分辅音时，用一摩擦音共振峰滤波器来模拟它。

共振峰合成器

- 共振峰合成的优点
  - 由于它是对声道的一种比较准确的模拟，因此可以合成自然度比较高的语音。
- 共振峰合成的缺点
  - 参数不好控制，从而对声道的模拟不精确，会影响合成语音的质量和自然度

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

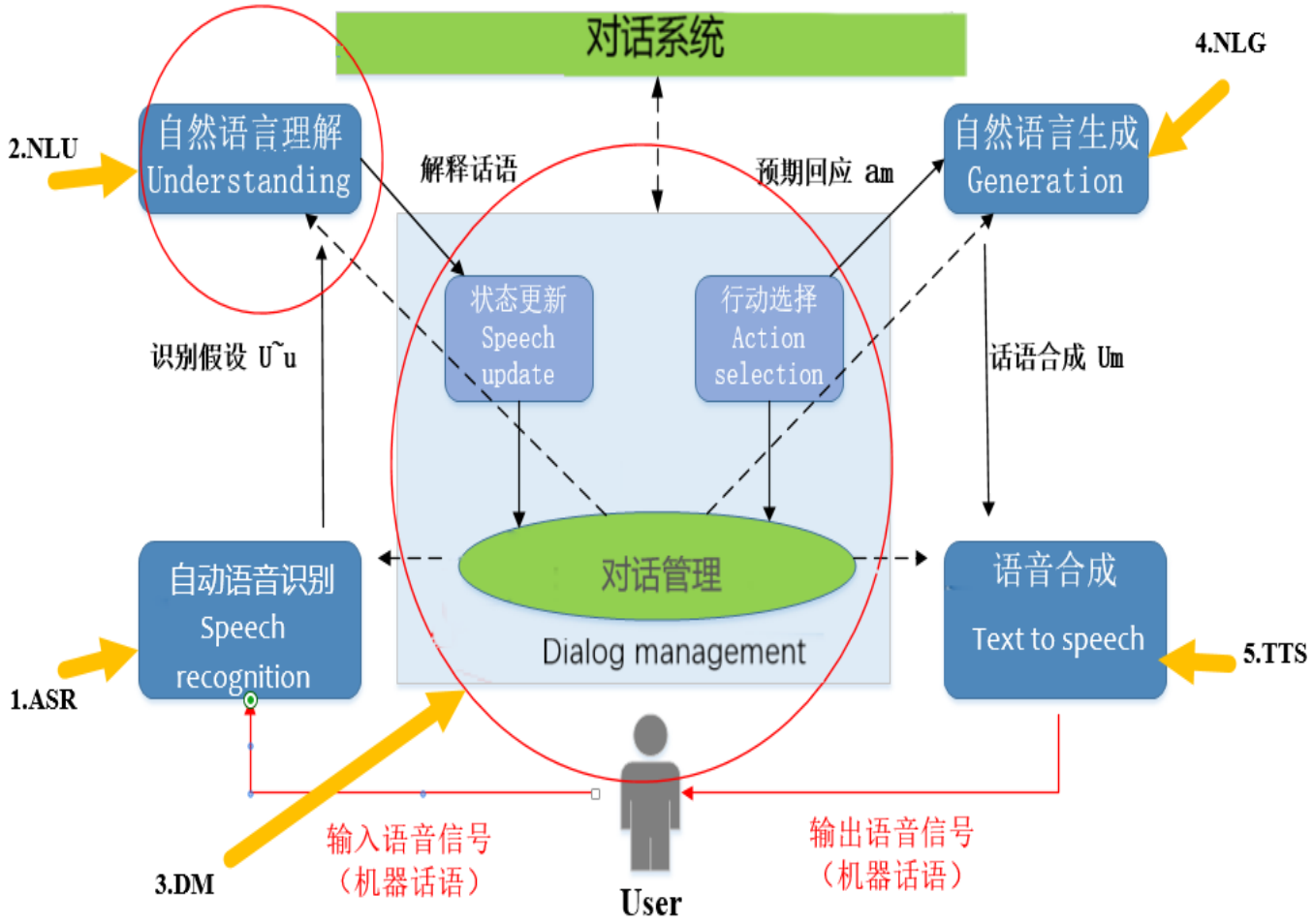
对话系统

- 对话系统（dialog system）是指以完成特定任务为主要目的的人机交互系统。在现有的人与人之间对话的场景下，对话系统能帮助提高效率、降低成本，比如客服与用户之间的对话。

例如

- Google：Siri
- 微软：Cortana
- 亚马逊的Echo

语音对话系统


对话系统
4.NLG
2.NLU
自然语言理解
Understanding
解释话语
预期回应 am
自然语言生成
Generation
状态更新
Speech
update
行动选择
Action
selection
识别假设 Ũu
话语合成 Um
对话管理
Dialog management
自动语音识别
Speech
recognition
语音合成
Text to speech
1.ASR
5.TTS
3.DM
输入语音信号
（机器话语）
输出语音信号
（机器话语）
User

自然语言理解

自然语言理解的目标是将文本信息转换为可被机器处理的语义表示。因为同样的意思有很多种不同的表达方式，对机器而言，理解一句话里每个词的确切含义并不重要，重要的是理解这句话表达的意思。为了让机器能够处理，我们用语义表示来表示自然语言的意思。语义表示可以用意图+槽位的方式来描述。意图即这句话所表达的含义，槽位即表达这个意图所需要的具体参数，用槽(slot)-值(value)对的方式表示。

自然语言理解

自然语言理解是所有对话系统的基础，目前有一些公司将自然语言理解作为一种云服务提供，方便其他产品快速的具备语义理解能力。比如Facebook的wit.ai、Google的api.ai和微软的luis.ai，都是类似的服务平台，使用者上传数据，平台根据数据训练出模型并提供接口供使用者调用。

对话管理

- 对话管理是对话系统的大脑，它主要干两件事情：
  - ➢维护和更新对话的状态。对话状态是一种机器能够处理的数据表征，包含所有可能会影响到接下来决策的信息，如对话管理模块的输出、用户的特征等；
  - ➢基于当前的对话状态，选择接下来合适的动作。

例：用户说「帮我叫一辆车回家」，此时对话状态包括对话管理模块的输出、用户的位置、历史行为等特征。在这个状态下，系统接下来的动作可能有几种：（1）向用户询问起点，如「请问从哪里出发」；（2）向用户确认起点，如「请问从公司出发吗」；（3）直接为用户叫车，「马上为你叫车从公司回家」。

对话管理方法

- 基于有限状态机

将对话过程看成是一个有限状态转移图。对话管理每次有新的输入时，对话状态都根据输入进行跳转。跳转到下一个状态后，都会有对应的动作被执行。基于有限状态机的对话管理，优点是简单易用，缺点是状态的定义以及每个状态下对应的动作都要靠人工设计，因此不适合复杂的场景。

对话管理方法

■ 部分可见的马尔可夫决策过程

所谓部分可见，是因为对话管理的输入是存在不确定性的，对话状态不再是特定的马尔可夫链中特定的状态，而是针对所有状态的概率分布。在每个状态下，系统执行某个动作都会有对应的回报。基于此，在每个对话状态下，选择下一步动作的策略即为选择期望回报最大的那个动作。

■ 优点：

（1）只需定义马尔可夫决策过程中的状态和动作，状态间的转移关系可以通过学习得到；

（2）使用强化学习可以在线学习出最优的动作选择策略。

对话管理方法

- 基于神经网络的深度学习方法

直接使用神经网络去学习动作选择的策略，即将对话理解的输出等其他特征都作为神经网络的输入，将动作选择作为神经网络的输出。这样做的好处是，对话状态直接被神经网络的隐向量所表征，不再需要人工去显式的定义对话状态。对话策略优化采用强化学习技术，决定系统要采取的行动指令。

自然语言生成

- 自然语言生成模块是根据对话管理模块输出的系统行动指令， 生成对应的自然语言回复并返回给用户。解决回复生成问题的方法目前主要分为两种，即基于检索的方法和基于生成的方法。
  - ➢检索式方法通常是在一个大的回复候选集中选出最适合的来回答用户提出的问题。
  - ➢生成式方法则借助于循环神经网络（RNN）通过对对话的学习来生成新的回复。

内容提要

- 听觉通路
- 听觉信息的中枢处理
- 语音编码
- 韵律认知
- 语音识别
- 语音合成
- 对话系统
- 言语行为

言语行为

奥斯汀（J. L. Austin）认为，说任何一句话时，人们同时要完成三种行为：言内行为、言外行为、言后行为。言外行为是通过一定的话语形式，通过协定的步骤与协定的力而取得效果，所以言外行为是协定的。而言后行为依赖于语境，不一定通过话语本身就能取得，因此是不确定的。由于“言内行为”属于语言体系的范围，“言后行为”本身又不是语言行动，而且听者的反应也不是一个语言过程，而是复杂的心理过程，所以语言学家过去不大讨论“言后行为”，而把注意力集中在“言外行为”上。

言语行为

奥斯汀（J.L.Austin）把言外行为分为五类：

- 判定语（verdictives）
- 裁定语（exercitives）
- 承诺语（commissives）
- 阐述语（expositives）
- 行为语（behabitives）

言语行为

塞尔（J.R.Searle）修改了这一分类，把言外行为分为“新五类”：

- 断言（assertives）
- 指令（directives）
- 承诺（commissives）
- 表达（expressives）
- 宣告（declarations）

言语行为

言语行为理论的提出，无论对语言研究还是对应用语言学、社会语言学、语用学以及语言习得的研究都产生了重大影响。一方面，它使学者们在有关方面的研究从以语法或语言形式为中心转向以言语功能为中心；从以单句为中心转向以语篇为中心；从以语言本身为中心转向以语言使用者、社团以及语言环境等为中心；另一方面，言语行为理论使诸多研究从以语言知识为中心转向以交际功能为中心。

思考题

- 5-1 请给出听觉基本神经通路模型。
- 5-2 听觉信息的中枢处理有哪些机制？
- 5-3 什么是语音编码？实现语音编码的方法有哪些？
- 5-4 为什么韵律在言语交流中起着非常重要的作用？
- 5-5 请画出语音识别系统的框图，并阐明各个模块的功能。
- 5-6 请概述常用的语音合成技术，并比较它们的优缺点。
- 5-7 什么是听觉场景分析？举例阐述场景分析的过程。

Thank You