Warm Up

- 文字输入方法
  - 键盘输入
    - 万码奔腾
  - 笔式输入
    - 联机手写

共性：用手输入！

其它方式？

第五讲 语音交互技术

提　纲

- 学科背景
- 语音识别概况
- 语音信号处理基础知识
- 语音识别技术概述
- 声学特征提取
- 典型识别技术简介
- 未来发展展望

语音信号处理

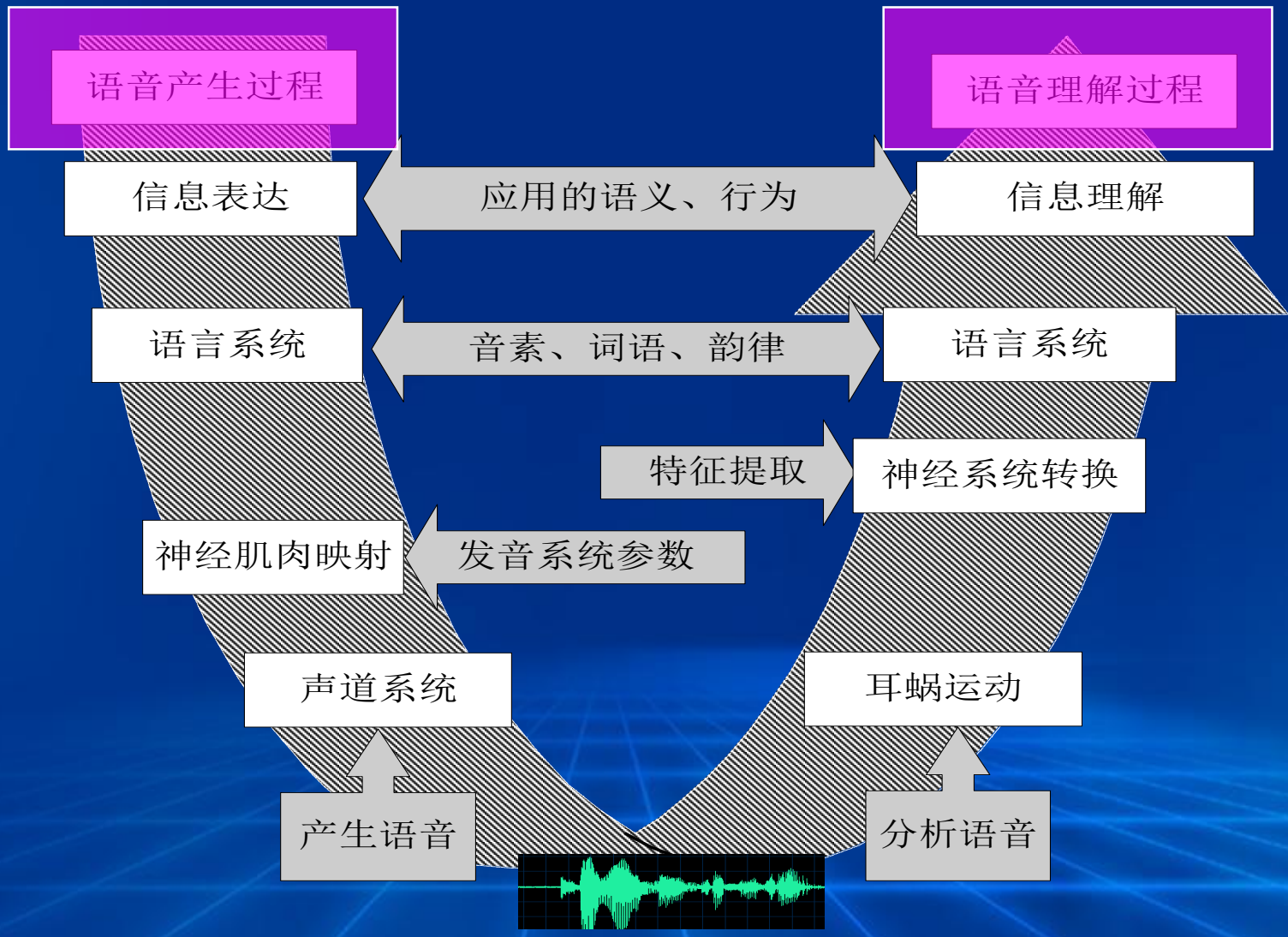
- 学科内涵
  - 利用数字信号处理技术对语音信号进行处理的一门学科，处理的目的是：
    - 得到某些参数以便高效传输或储存
    - 进行人工合成语音、辨识讲话人、识别出讲话内容、语音增强等应用
    - 实现自然的人机对话交流

语音信号处理与人类功能类比

- 聋、哑、盲→耳聪目明，听说读写

- 语音交互涉及两类器官
  - 发声器官——语音合成
    - 肺、气管、喉（包括声带）、咽、鼻、口
  - 耳朵——语音分析、识别、话者识别
    - 外耳、中耳、内耳

语音产生语音理解生理过程



相关研究方向

- 语音压缩
  - 高效存储、传输语音信号
- 语音合成
  - 输出自然可懂的语音信号
- 语音识别
  - 提取或匹配语义
- 话者识别
  - 安全应用
- 语音增强
  - 提高信噪比、加重语音成分

语音信号处理学科特点

- 多学科交叉特性
  - 计算机学科 ——计算机智能接口
  - 信息处理学科——信息识别及提取
  - 人工智能——时序模式、多维模式识别
  - 通信及电子系统——信源处理
  - ……
  - 语音学、声学、语言学、认知科学、生理学、心理学、数理统计等多学科交叉

语音信号处理发展简况

- 1960年代
  - 特别是数字信号处理方法和算法，数字滤波器，FFT
- 1970年代
  - 用于语音信号的信息压缩和特征提取的线性预测技术(LPC)，用于语音分析、合成等
  - 输入语音与参考样本间的动态时间规正DTW技术
- 1980年代
  - 矢量量化(VQ)：基于聚类分析的高效数据压缩技术
  - HMM：描述语音信号产生过程
  - ANN：人工神经网络
- 1990年代
  - 语音识别、语音合成技术的重要发展与应用阶段

语音识别概况

语音识别应用价值

- 信息查询（股票、天气、航班……）
- 人机界面（新一代操作系统、智能家居）
- 听写机（文字输入、记录） ……
- 数据检索管理（语音数据检索）……
- 医疗教育生活（语音控制交互）……
- 智能机器人（人机对话，命令） ……
- ……

语音识别系统分类及当前技术特点

技术分类参数	参数变迁情况
说话模式	孤立词→连续语音
说话风格	朗读语言→口语
话者	说话人有关→说话人无关
词汇量	小词汇量(数10) →大词汇量(20000)
语言模型	有限句型文法→上下文无关文法
信噪比(SNR)	30db→10db
传输通道	麦克风→电话语音

典型技术类型与应用

- 大词汇量连续语音识别系统LVCSR
  - 听写机， IBM ViaVoice 听写机
- 受限词汇量基于语法的ASR
  - 旅馆预定， AT&T VRCP系统（自助话务员协助呼叫）， NTT ANSER 语音识别银行服务系统
- 连续语流中的关键词识别
  - 语音检索， SONY AIBO 机器狗
- 孤立词识别
  - 手机Name Dialing

语音识别历史发展

- 50年代
  - AT&T Bell Lab，可识别10个英文数字
- 60年代
  - LP较好地解决了语音信号产生模型, DP则有效解决了不等长语音的匹配问题。
- 70年代
  - DTW技术基本成熟，VQ和HMM理论；实现了基于LPC和DTW技术相结合的特定人孤立语音识别系统。
- 80年代
  - HMM模型和ANN在语音识别中成功应用，1988年美国CMU大学基于VQ/HMM开发SI-CSR系统 SPHINX。
- 90年代一大规模应用，工业标准，各行各业逐渐实用。
- 21世纪—与人工智能深度结合，语音智能的广泛应用。

语音识别发展迅速的原因

- 坚实的理论基础以及算法
  - 以HMM为代表的统计方法
  - 以ANN为代表的神经网络方法
- 大规模语音、文本数据库的建立
- 标准的测试和评估体系
- 计算机硬软件的快速发展

Math is King, data is queen!

标准是催化剂！

技术现状

- 语音识别在受限条件下已取得重大进展，在技术上相对比较成熟
  - 例如办公环境下具有说话人自适应能力的专用文本标准语音听写识别，恶劣环境下的专用小词汇量识别，小词汇量非特定人命令识别；小词汇量关键词检测；专用领域的中、小词汇人机对话等等；
- 除了一些通用环境下的语音识别外，几乎所有的系统需要针对某一个应用需要进行工程化的设计和实现
  - 对于不同的应用解决问题的着重点不一样，而目前的语音识别还不能在同一框架下去解决所有复杂的问题，因而有针对性地解决问题的方法至关重要。

技术挑战

- 数据资源 (年龄、性别、语言、方言、主题、情绪、地域……切分、标注体系)
- 抗噪性能(背景噪声、信道噪声、干扰)
- 协同发音(Co-articulation)
  - Seat and Soup
- 口语现象(重复、顿措、语序颠倒……)
- 说话人变异(口音、情绪、年龄……)
- 听觉机理(音量、频率、抗噪、区分……)

语音识别的性能评价

原句：我 们 明 天 去 天 安 门

识别：我 × 明后天 去 天 坛 ×

删除错误 Deletion　　插入错误 Insertion　　替换错误 Substitution

正确率：$$Correct = \frac{N - D - S}{N} \times 100\%$$

准确率：$$Accuracy = \frac{N - D - S - I}{N} \times 100\%$$

准确率 <= 正确率：系统正确率很高，其准确率未必高！

语音信号处理基础知识

语音和语言

- 声音与语音
  - 语音
    - 人讲话所发出的声音
    - 由一连串的“音”排列组成，是包含语言的声音
  - 音的排列规则及其含义→语言学研究范畴
  - 音的分类和研究→称为语音学

语言/语音的结构

- 篇章（一次演讲）
- 段落
- 句子
- 单词（词）
  - 语言的最小**语义**单元
- 音节
- 音素
  - 语言的元素，最小基本单元
  - 如：英语共有48个音素

音素

- 分类
  - 元音和辅音
- 元音
  - 发音时声带振动，呼出的气流通过口腔时不受阻碍，这样形成的语音称作元音。
  - 元音发音响亮，口腔中气流不受阻碍，是构成音节的主要音
    - 已知语言中最少2个，最多12个
    - 汉语单元音：a, o, e, i, u, v；还有双元音
- 辅音
  - 不论声带振动与否，发声时呼出的气流通过口腔或鼻腔时受到一定阻碍，这样的语音称为辅音。
  - 辅音发音不响亮，口腔中气流受到阻碍，不是构成音节的主要音

音节

- 音素的组合形成音节
  - 每个音节可以是一个元音和1~2个辅音组合
- 并非所有音素的组合都会成为音节
  - 例如：汉语里面dv
  - 意味着：音节数远少于音素的组合数

其他

- 重音
  - 西方语言的重要特点
- 语调
  - 讲话声音的调节
- 声调
  - 1，2，3，4

汉语语言/语音学特点

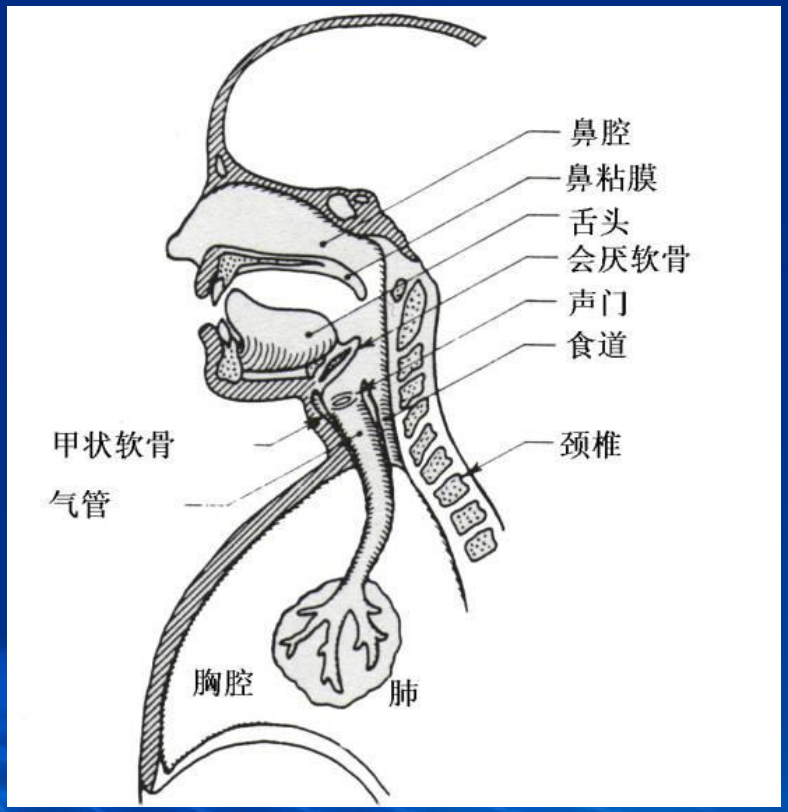
- 也有元音和辅音之分
- 声母和韵母之分更常用
  - 声母：汉字音节开头的辅音
  - 韵母：汉字音节除了声母之外的部分
  - 21个声母，39个韵母
- 汉语的自然单位是音节
  - 每个字都是单音节字
  - 一个音节一个字音
  - 字是独立的发音单位
- 声调
  - 阴平、阳平、上声、去声

汉语语言/语音学特点

- 音素少、音节少
  - 64个音素
  - 400个左右的音节
    - 400个左右的基本发音
    - 加上音调，1200多个不同的发音
    - 覆盖数万个汉字

语音产生及其声学特性

- 语音产生器官
  - 发声器官——语音合成
    - 肺、气管、喉（包括声带）、咽、鼻、口。它们共同形成一条形状复杂的管道
  - 声带和声门
    - 喉与气管的接口处称为声门
  - 声道
    - 其中喉以上的部分称为声道，随着发出语音的不同其形状是变化的



语音产生过程

- 产生语音的能量，绝大多数来源于正常呼吸时肺部呼出的稳定气流
  - 有极少数语种，如某些非洲语言，是利用吸气气流来发音的
- 声带——最重要的发声器官
  - 既是一个阀门又是一个振动部件
  - 呼吸时左右两声带打开（声门开）
  - 在说话的时候合拢，肺部气流经气管形成冲击“打开-闭合-打开-闭合-…”声门，从而冲击声带产生振动，然后通过声道响应变成语音

语音的频率性质

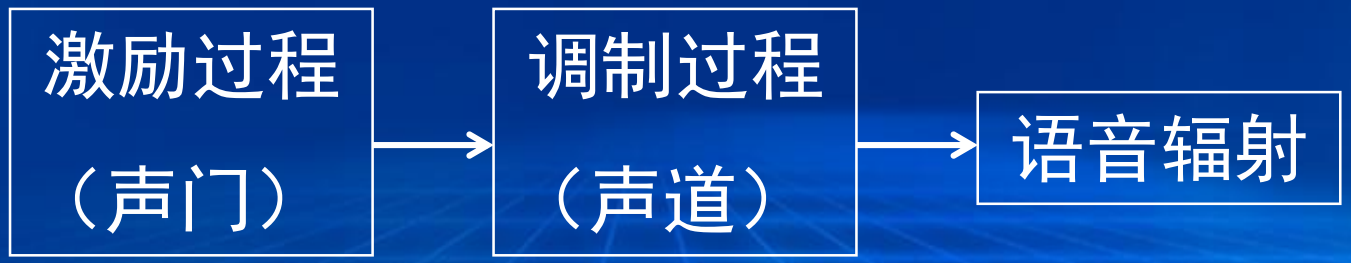
- 音调周期/基音周期
  - 声门开启-闭合一次的时间即振动周期
- 基频
  - 基音周期的倒数，声带振动的基本频率
- 音调
  - 声带振动的频率（即基音）决定了声音频率的高低，频率快则音调高，否则音调低
  - 人的基音范围
    - 70~350HZ，儿童和青年女性偏高，男性偏低

声道调制

- 声道
  - 咽、口腔和鼻腔
  - 从声门延伸至口唇的非均匀截面的声管，约17cm
- 功能
  - 谐振腔：放大某一频率而衰减其他频率分量
    - 谐振频率：由每一瞬间的声道外形决定，又称为共振峰，是声道的重要声学特征

发声过程小结

- 人的发声过程包括两个步骤
  - 声门/声带产生不同频率的声音
    - 准周期气流脉冲或白噪声
  - 声道对声源的调制作用



语音识别技术概述

语音传输信道描述

- **M: Message**
- **W: Word**
- **S: Speech**
- **A: Acoustic**
- **X: Signal/Feature**

基于词转换范式的语音识别原理

- 输入：声学语音信号序列X
- 输出：词序列（最小语义单元序列）
- 方法：贝叶斯模型范式

目标为(MAP)：$W = \arg\max_{\Omega} P_{\Omega}(W / X)$

后验概率为：

$P_{\Omega}(W / X) = P_{\Omega_X}(X / W) P_{\Omega_W}(W) / P(X),$

其中：

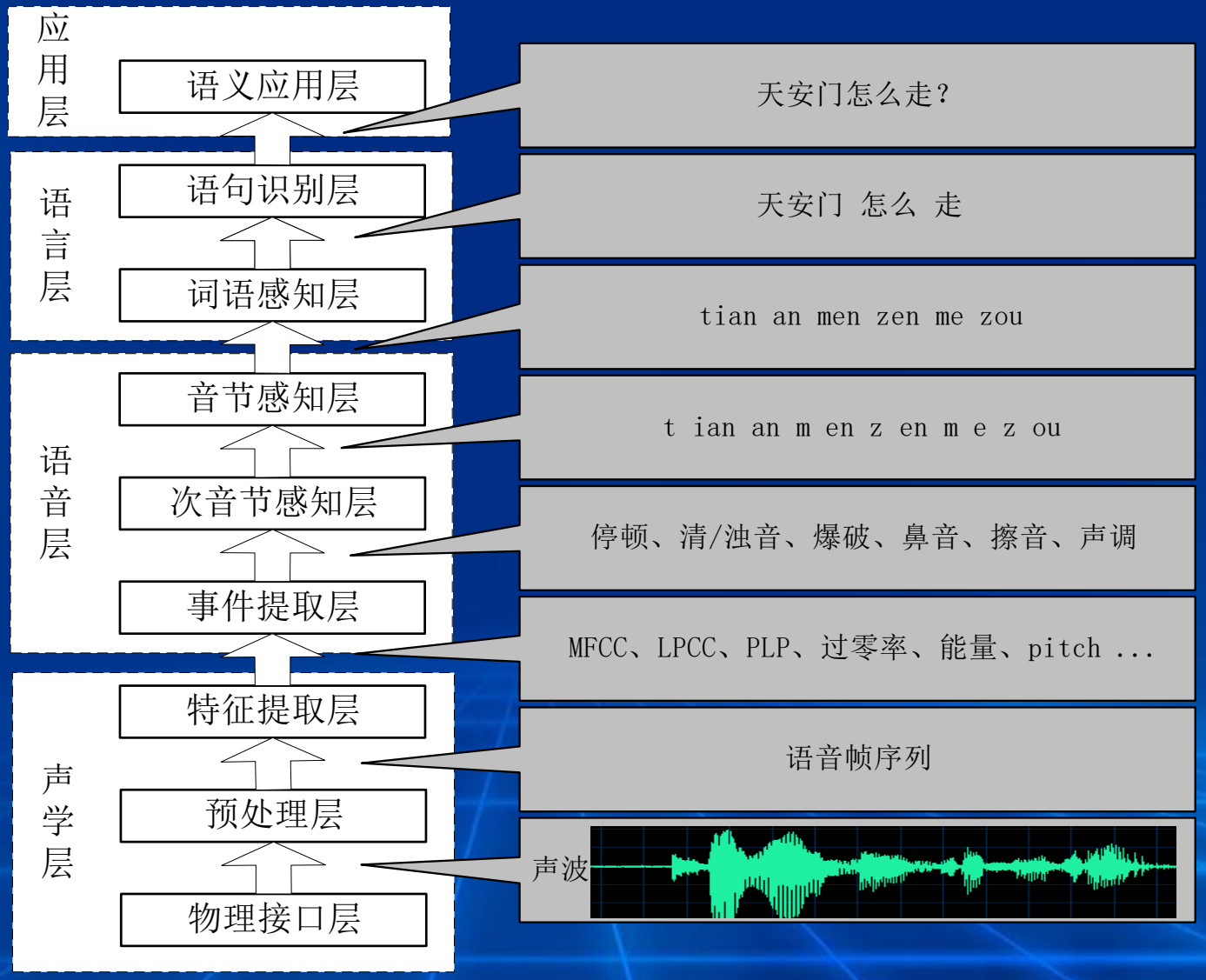
$P_{\Omega_X}(X / W)$表示声学模型给出的类条件概率密度,

$P_{\Omega_W}(W)$表示语言模型给出的W的先验概率

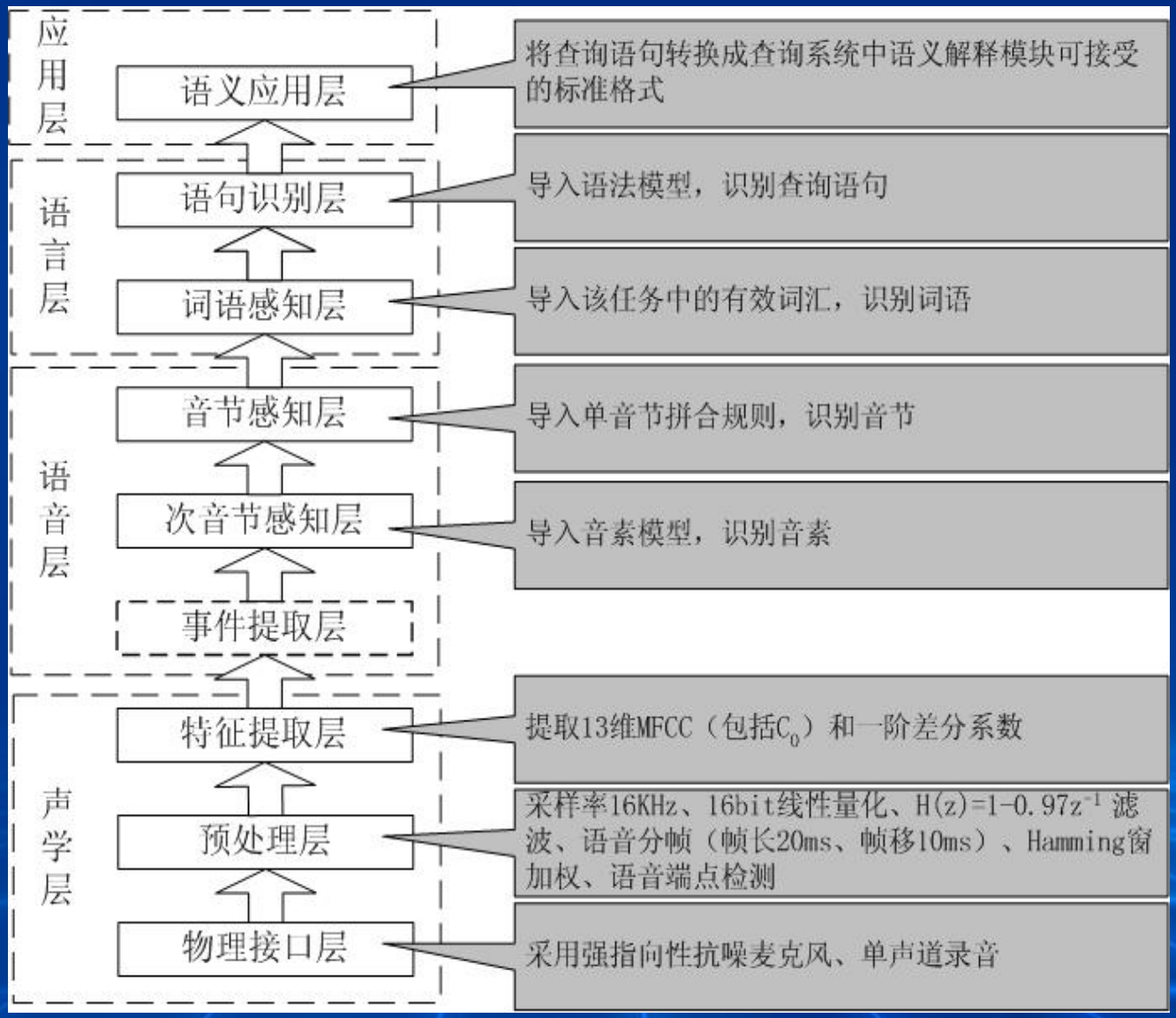
基于词转换范式的语音识别原理

- 声学模型
  - HMM
  - GMM
  - ANN
- 语言模型
  - N-Gram
  - Rule-based

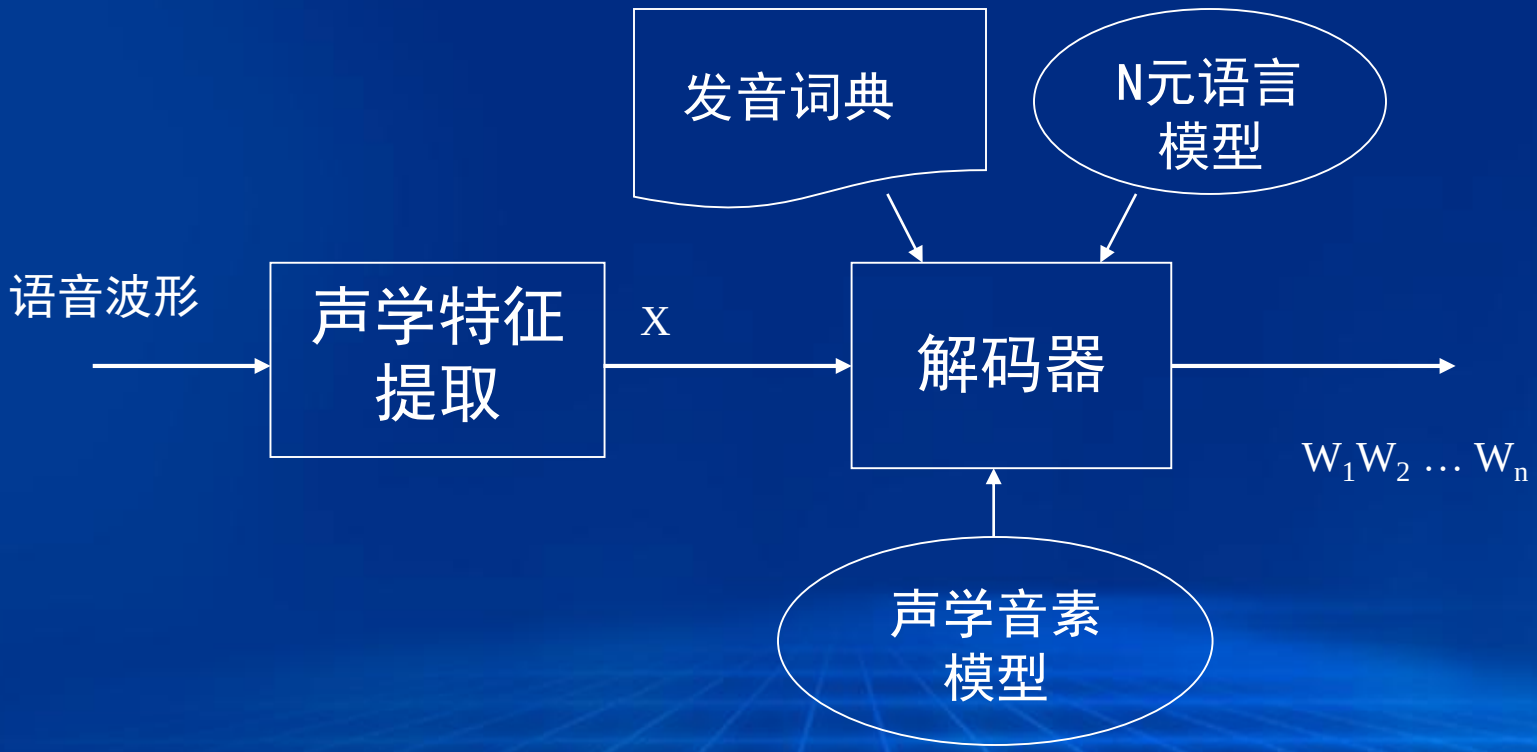
语音识别层次模型


应用层
语义应用层
语言层
语句识别层
词语感知层
语音层
音节感知层
次音节感知层
事件提取层
声学层
特征提取层
预处理层
物理接口层
天安门怎么走？
天安门 怎么 走
tian an men zen me zou
t ian an m en z en m e z ou
停顿、清/浊音、爆破、鼻音、擦音、声调
MFCC、LPCC、PLP、过零率、能量、pitch ...
语音帧序列
声波

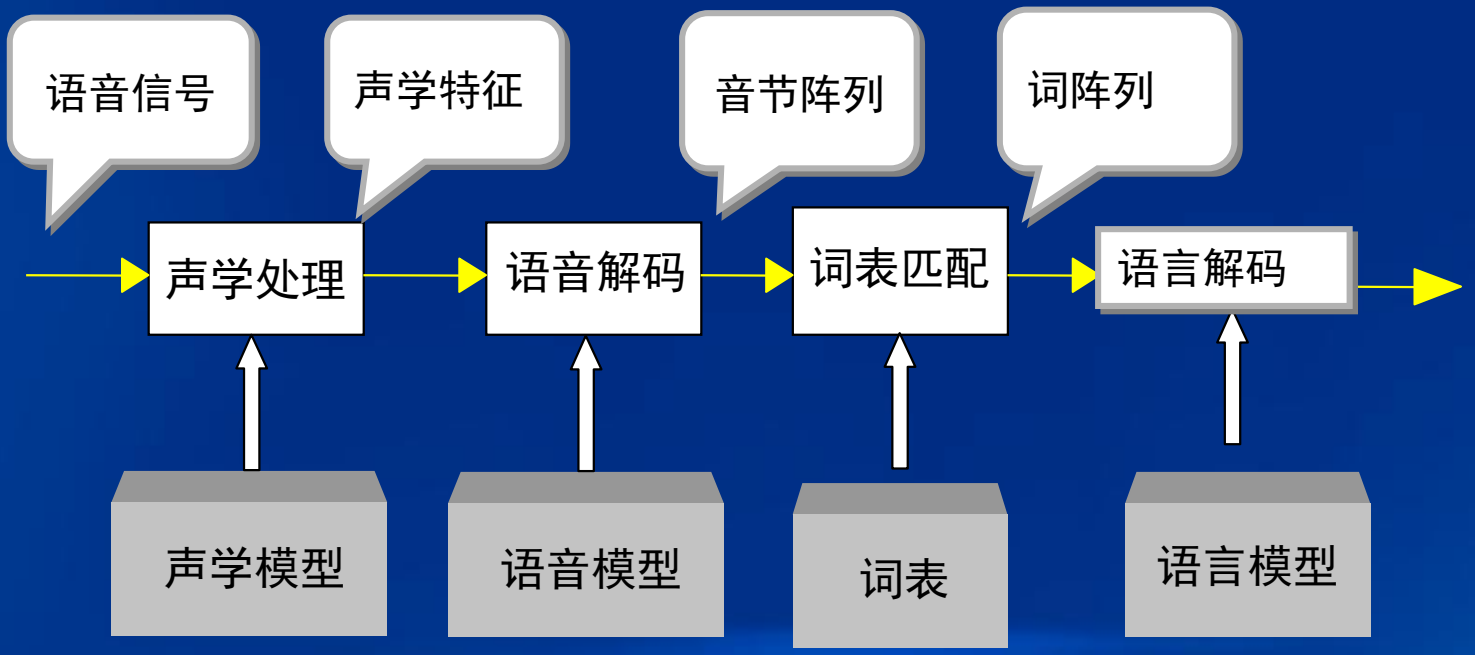
层次模型—系统设计


应用层
语义应用层
将查询语句转换成查询系统中语义解释模块可接受的标准格式
语言层
语句识别层
导入语法模型，识别查询语句
词语感知层
导入该任务中的有效词汇，识别词语
语音层
音节感知层
导入单音节拼合规则，识别音节
次音节感知层
导入音素模型，识别音素
事件提取层
声学层
特征提取层
提取13维MFCC（包括C_0）和一阶差分系数
预处理层
采样率16KHz、16bit线性量化、$H(z)=1-0.97z^{-1}$ 滤波、语音分帧（帧长20ms、帧移10ms）、Hamming窗加权、语音端点检测
物理接口层
采用强指向性抗噪麦克风、单声道录音

语音识别框架

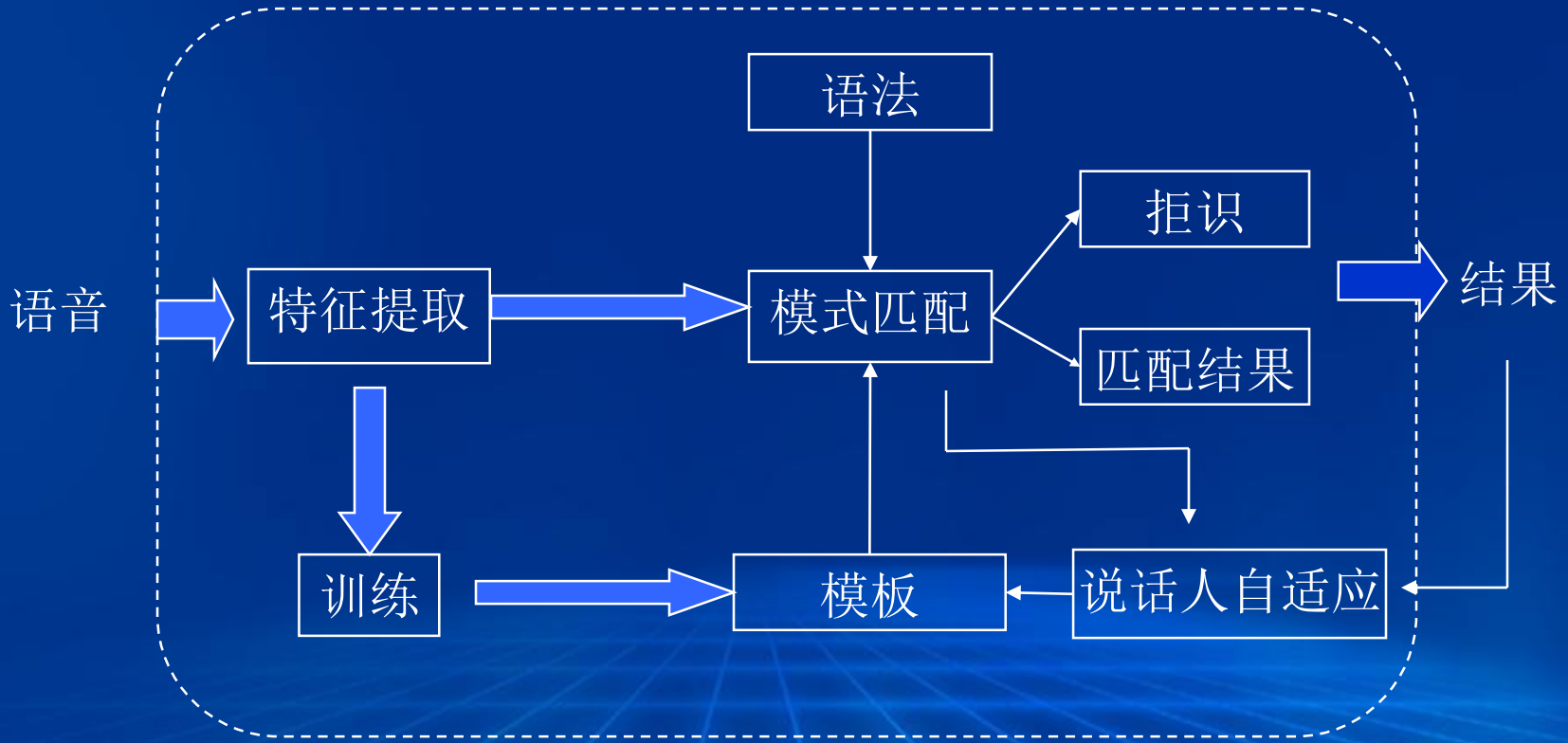


语音识别原理图



对于不同类型的系统，并不需要所有上述模块
例如：在小词汇表命令词识别系统中

语音识别系统基本构成



典型统计语音识别方法

- 模板匹配法
- 随机模型法
  - HMM
- 概率语法分析法
- 神经网络方法

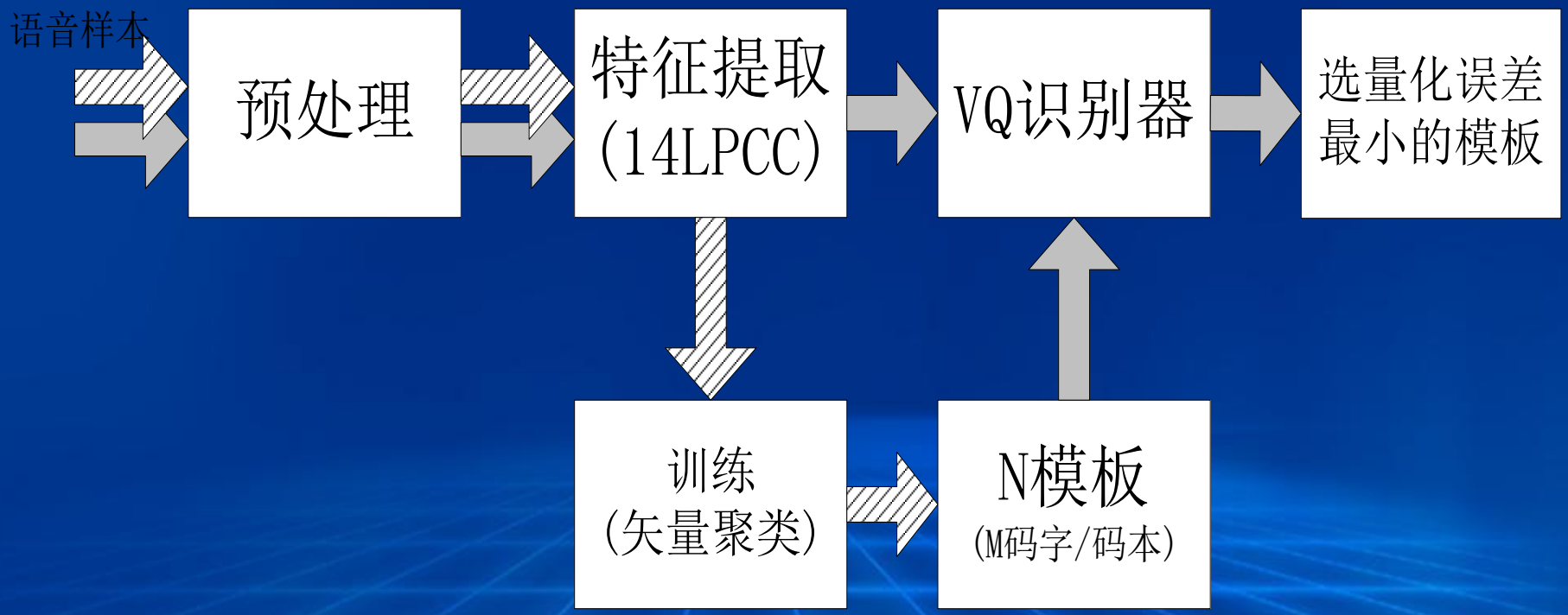
模板匹配法

- 将测试语音与模板的参数一一比较
  - 特定人、小词汇、孤立人识别系统
  - 动态规划，动态时间规正DTW
- 判决依据
  - 失真度最小准则
  - 距离最小准则
  - 相似度最大准则
- 难以实现鲁棒
  - 语速，讲话人，噪声…

随机模型法

- 利用HMM概率参数来对似然函数进行估计与判决，从而得到识别结果的方法
- 通过HMM的状态函数，较好的利用了语言结构的动态特性

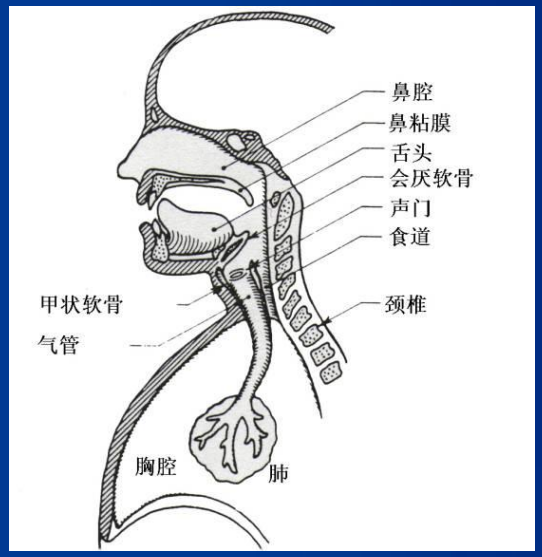
语音识别系统举例



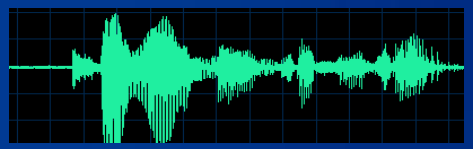
典型成功系统

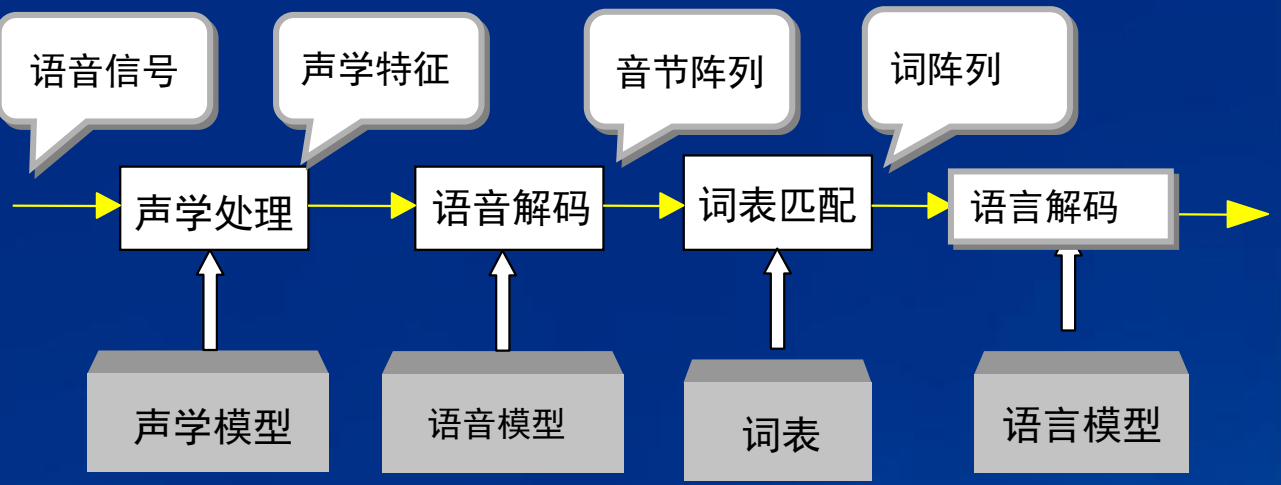
- 1990年代以来
  - 大词汇量，非特定人，连续语音识别
- 声学特征
  - LPCC， MPCC，LPLCC
- 识别方法
  - HMM为统一框架
  - 为每个基本识别单元建立至少一套HMM结构和参数

声学特征提取



激励过程（声门） → 调制过程（声道） → 语音辐射





- 声学特征
  - 时域特征
  - 频域特征
  - 听觉特征

时域特征

- 短时平均能量

$$Eng(t)=\frac{1}{N}\sum_{n=0}^{N-1}S_t^2(n)$$

  - N为分析窗的宽度，$S_t(n)$ 表示第t 帧中第n个采样点的信号值
- 短时平均过零率

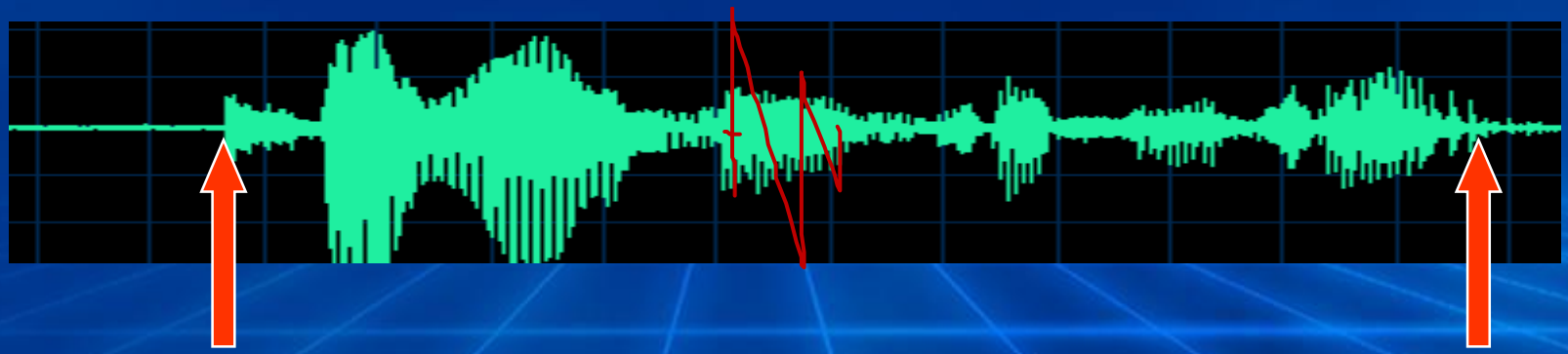
$$ZCT(t)=\frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{2}\left[1-Sgn\left(S_t(n)\cdot S_t(n-1)\right)\right]$$

其中符号函数定义为

$$Sgn(x)=\begin{cases}1, & x>0\\ -1, & x<0\end{cases}$$

时域特征

- 能量和过零率参数的缺点
  - 对于说话人和背景噪声的鲁棒性较差
- 目前时域参数多用在语音的预处理上
  - 如端点检测，判断语音的开始与结束
  - 也有人把它作为模型参数进行使用

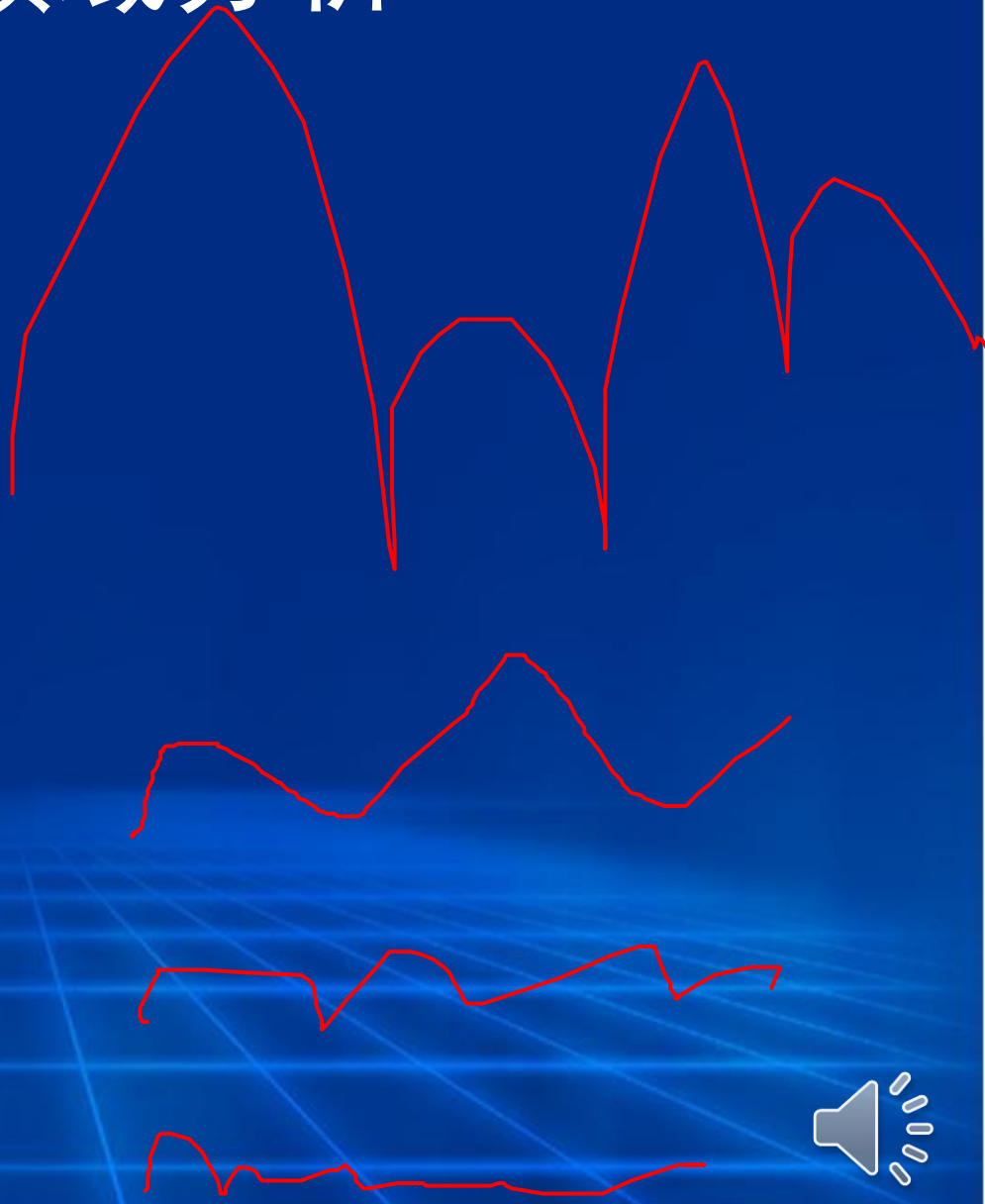


频域分析

- 为什么要进行频域分析？
  - 稳态语音的生成模型由线性系统组成，其被一随时间做周期变化或者随机变化的源所激励，因而系统输出频谱反映了激励与声道频率响应特性
  - 语音信号的频谱具有非常明显的语音、声学意义，可以获得重要的语音特征，如共振峰
    - 共振峰是指在声音的频谱中能量相对集中的一些区域
    - 共振峰不但是音质的决定因素，而且反映了声道（共振腔）的物理特征。
      - 声音在经过共振腔时，受到腔体的滤波作用，使得频域中不同频率的能量重新分配，一部分因为共振腔的共振作用得到强化，另一部分则受到衰减，得到强化的那些频率在时频分析的语图上表现为浓重的黑色条纹。由于能量分布不均匀，强的部分犹如山峰一般，故而称之为共振峰。
    - 在语音声学中，共振峰决定着元音的音质

广义频域分析

- 广义频谱分析
  - 频谱
  - 倒频谱
  - 功率谱
  - 频谱包络
- 常用频谱分析方法
  - 傅里叶变换法
  - 线性预测法
  - 带通滤波器组法



频域分析

- 基本工具——傅里叶变换
  - 标准傅里叶变换
    - 适用于周期、平稳随机信号
    - 不适合于非平稳的语音信号
- 短时傅里叶变换
  - 短时谱，有限长度的傅里叶变换
    - 即对某一帧语音进行傅里叶变换
    - 窗选语音信号的标准傅里叶变换
  - 特别适用于“语音分析和语音合成”
    - 因为其可以精确的恢复语音波形

声学特征小结

- 短时平均能量/幅度/功率
- 短时平均过零率
- 线性预测系数（LPC）
- LPC倒谱特征（LPCC）
- Mel 倒谱参数（MFCC）

识别方法

- 对一个语音波形序列，经过短时分帧特征提取，得到特征矢量序列

$$Y=\{y_1,y_2,\cdots,y_N\}$$

- 问题提出
  - 如何对其建模？
  - 如何将其与已经建成的模型比对？
  - 长度不一的特征序列又如何去时间对准？

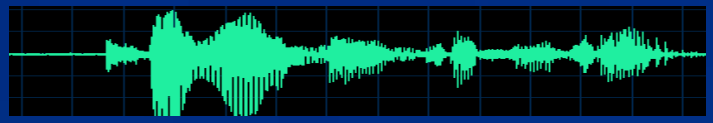
语音识别常用模式匹配方法

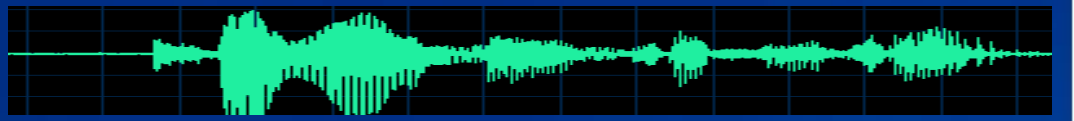
- 动态时间规整(DTW)
- 矢量量化(VQ)
- 隐马尔科夫模型(HMM)
- 时延神经网络(TDNN)
- 模糊逻辑算法

动态时间规整（DTW）

动态时间规整

- 语音识别模式匹配的问题——时间对准
  - 同一个人在不同时刻说同一句话、发同一个音，也不可能具有完全相同的时间长度
  - 语音的持续时间随机改变，相对时长也随机改变
- 方法1：线性时间规整

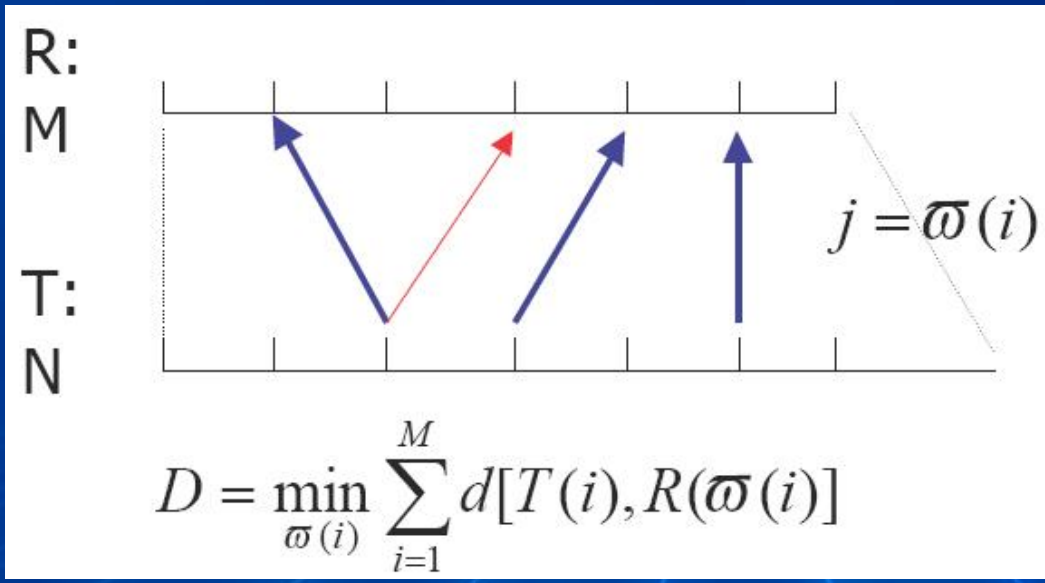


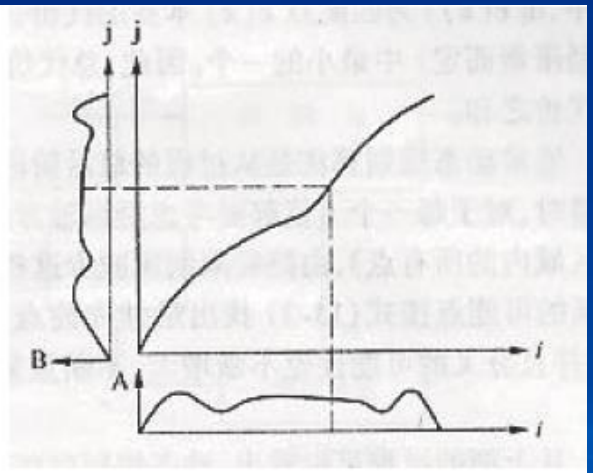


  - 均匀伸长或缩短
  - 依赖于端点检测
    - 通过时域分析进行，利用能量、振幅和过零率等特征
  - 缺点：仅扩展时间轴，无法精确对准
- 方法2：动态时间规整
  - DTW－Dynamic Time Warping

DTW的基本思想

- 一种非线性时间规整模式匹配算法
  - 将时间规整与距离测度结合起来，采用优化技术，以最优匹配为目标，寻找最优的时间规整函数w(i)，从而实现大小(长短)不同的模式的比较



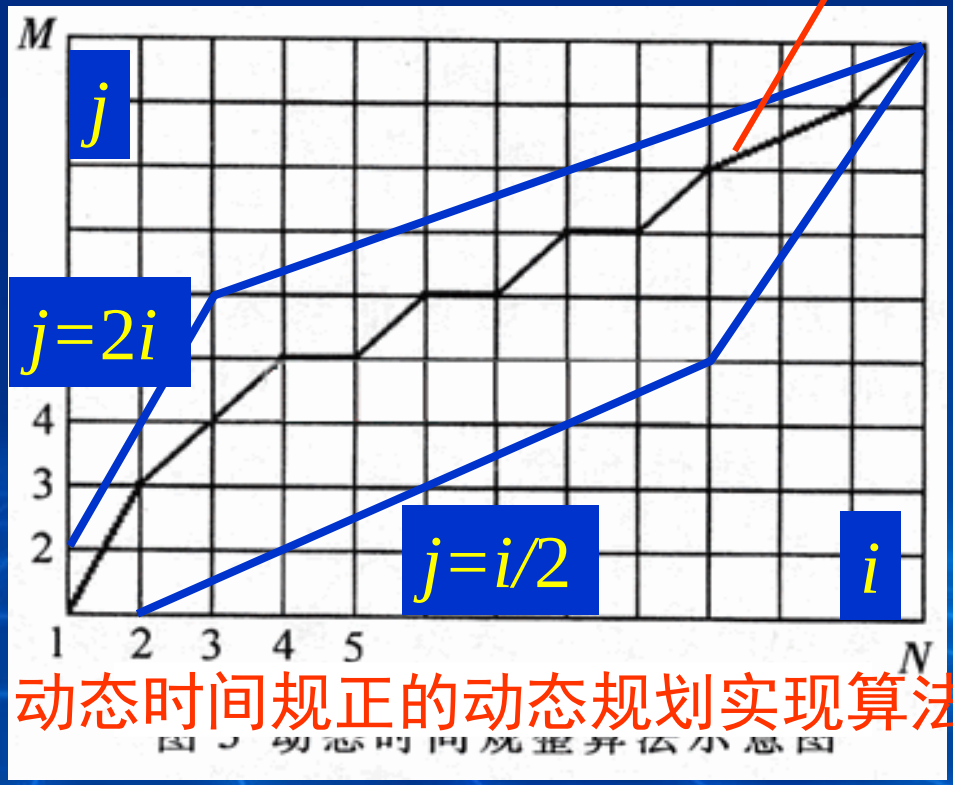


DTW的DP实现

- 动态规划

$$D[c(k)]=d[c(k)] + \min D[c(k-1)]$$

  - 搜索区域约束
    - 平行四边形
      - j=2i
      - j=i/2
  - 路径限制
    - W斜率
      - 0，1，2



动态时间规正的动态规划实现算法

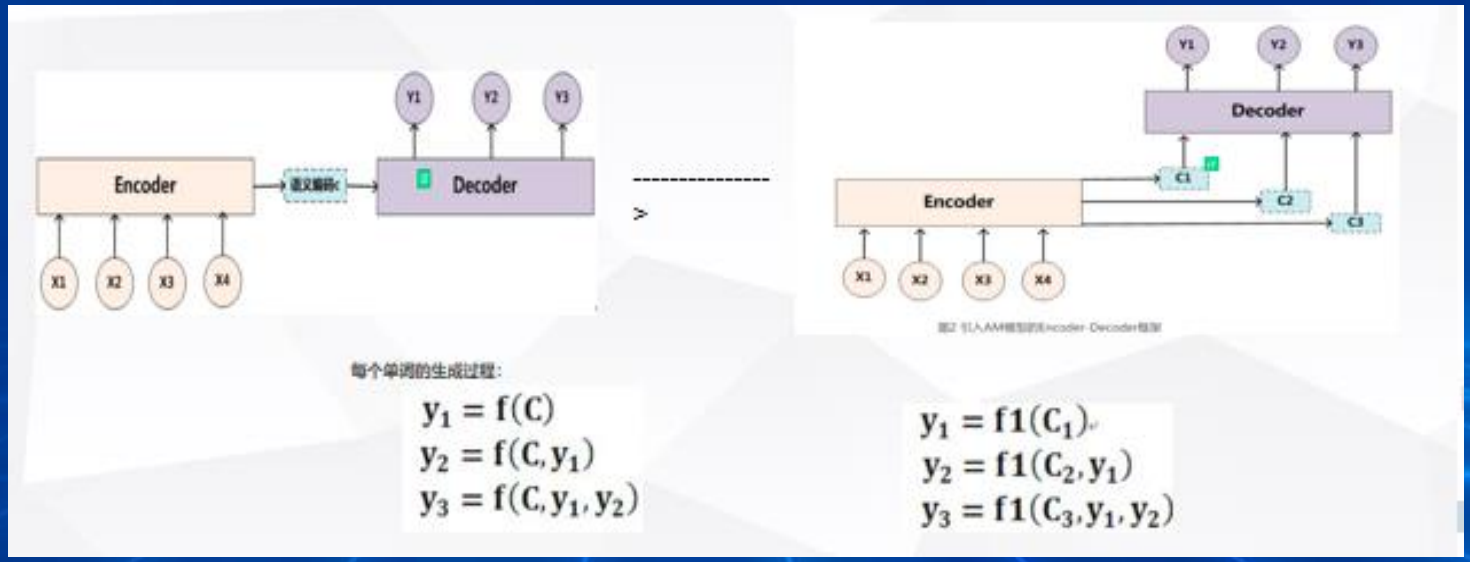
DTW评价

- 适用场合
  - DTW适合于特定人、基元较少的场合
  - 多用于孤立词识别
- DTW的问题：
  - 运算量较大；
  - 识别性能过分依赖于端点检测；
  - 太依赖于说话人的原来发音；
  - 不能对样本作动态训练；
  - 没有充分利用语音信号的时序动态特性；

端到端的语音识别

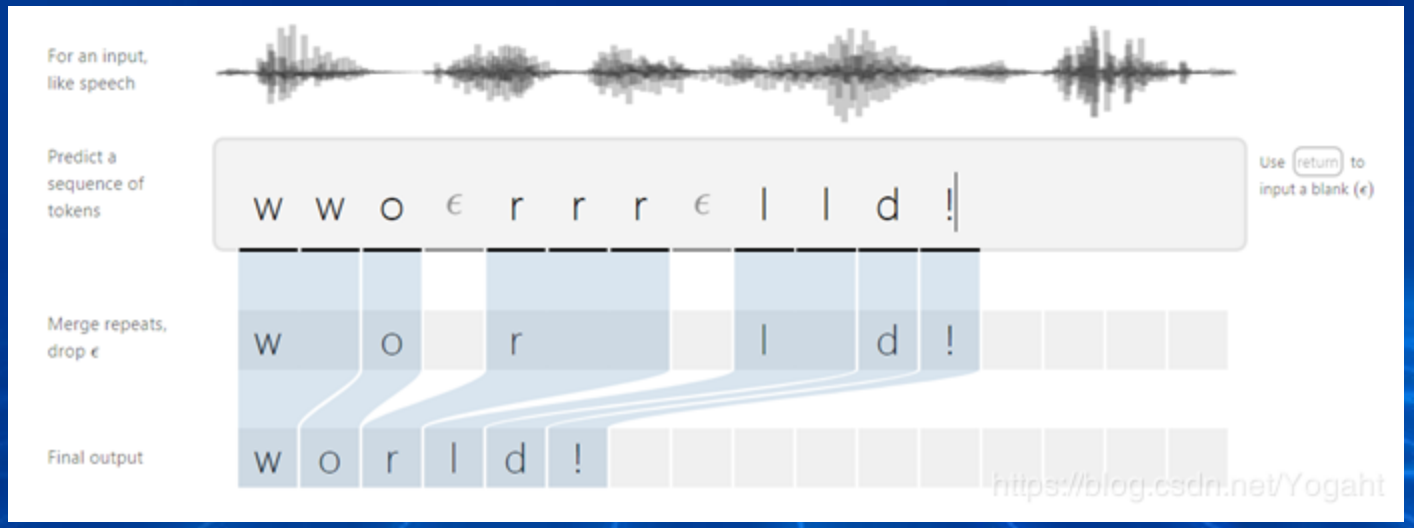
端到端的语音识别

- 直接从声学特征输入就可以得到识别的词序列，去掉HMM
  - CTC模型（Connectionist Temporal Classification）
  - 基于Attention的Encoder-decoder模型



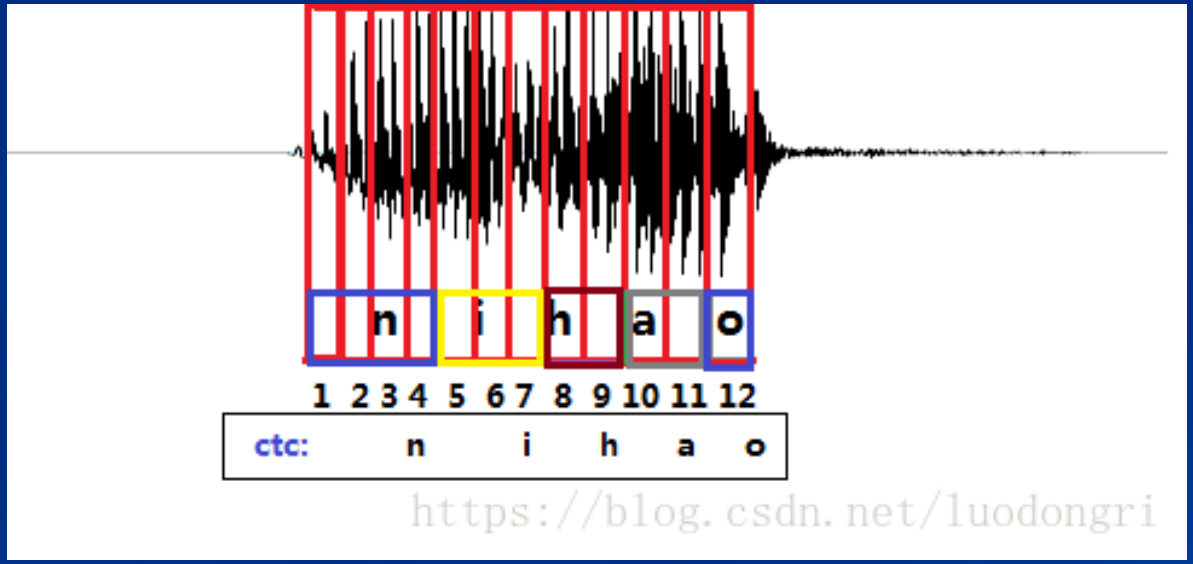
CTC模型

- Connectionist Temporal Classification，连接时序分类，是一种用于序列建模的工具，其核心是定义了特殊的目标函数/优化准则。
- CTC算法将网络模型输出和标签进行对齐。



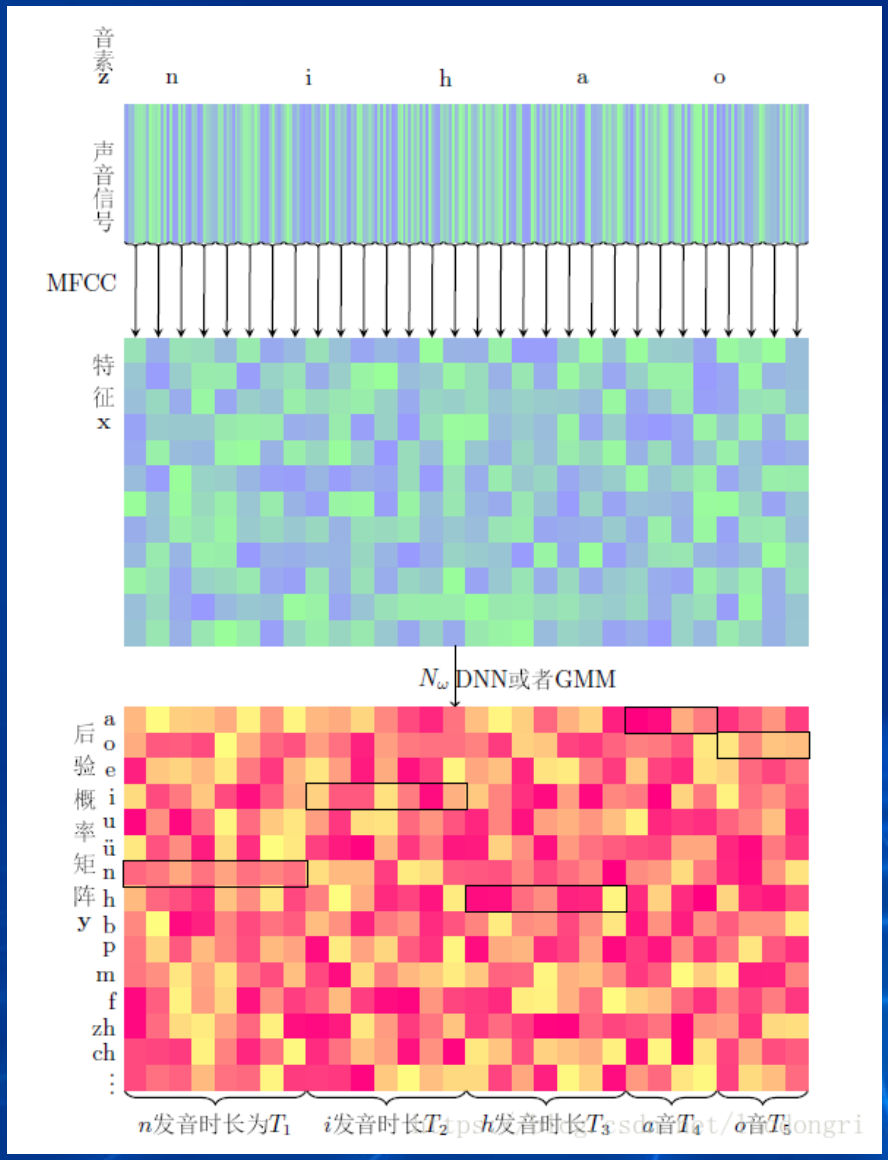
若不进行对齐，输出为“wworrrlld！”

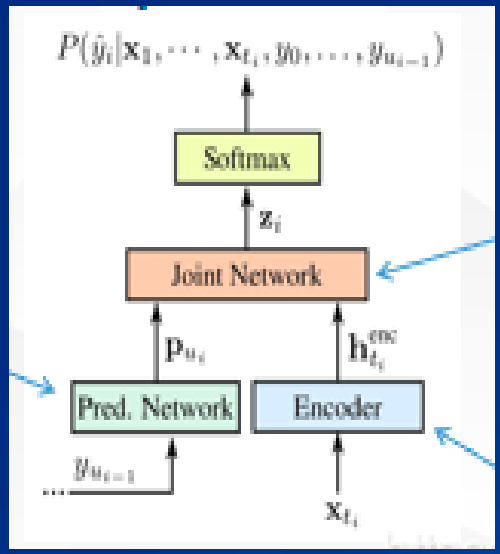
CTC模型



CTC引入了blank（该帧没有预测值），每个预测的分类对应的一整段语音中的一个spike（尖峰），其他不是尖峰的位置认为是blank。

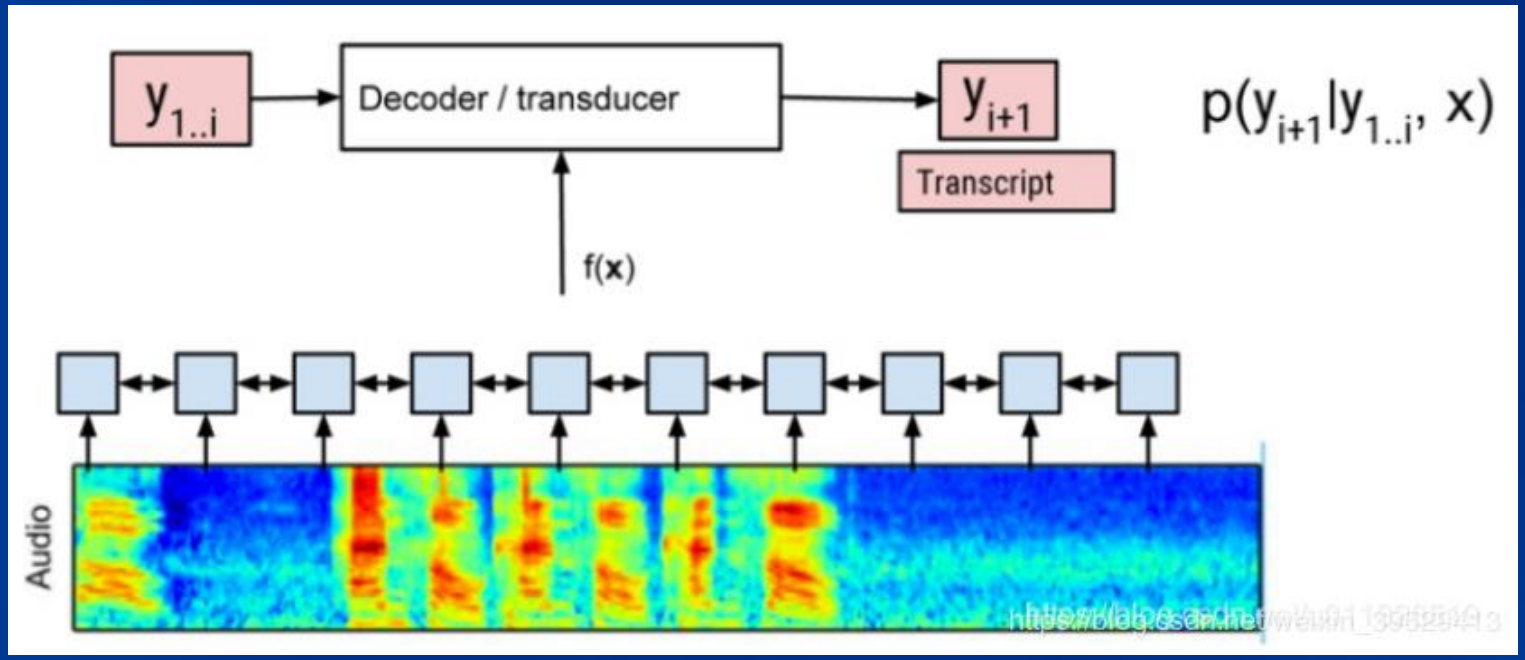
RNN+CTC模型训练



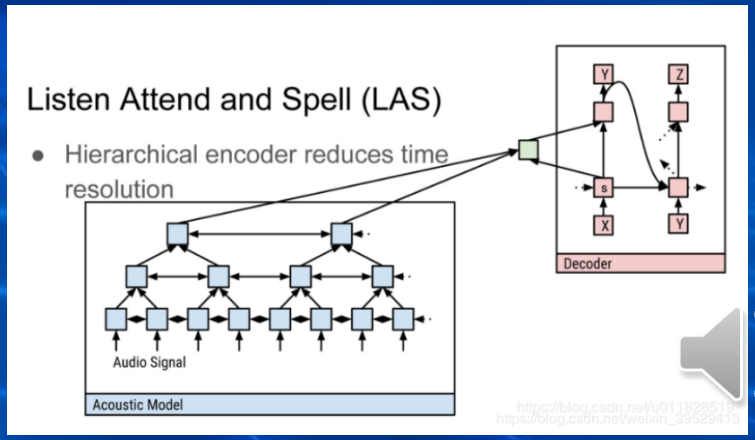


在CTC模型的Encoder基础上，又加了一个输出作为输入的一个RNN，称为预测网络。

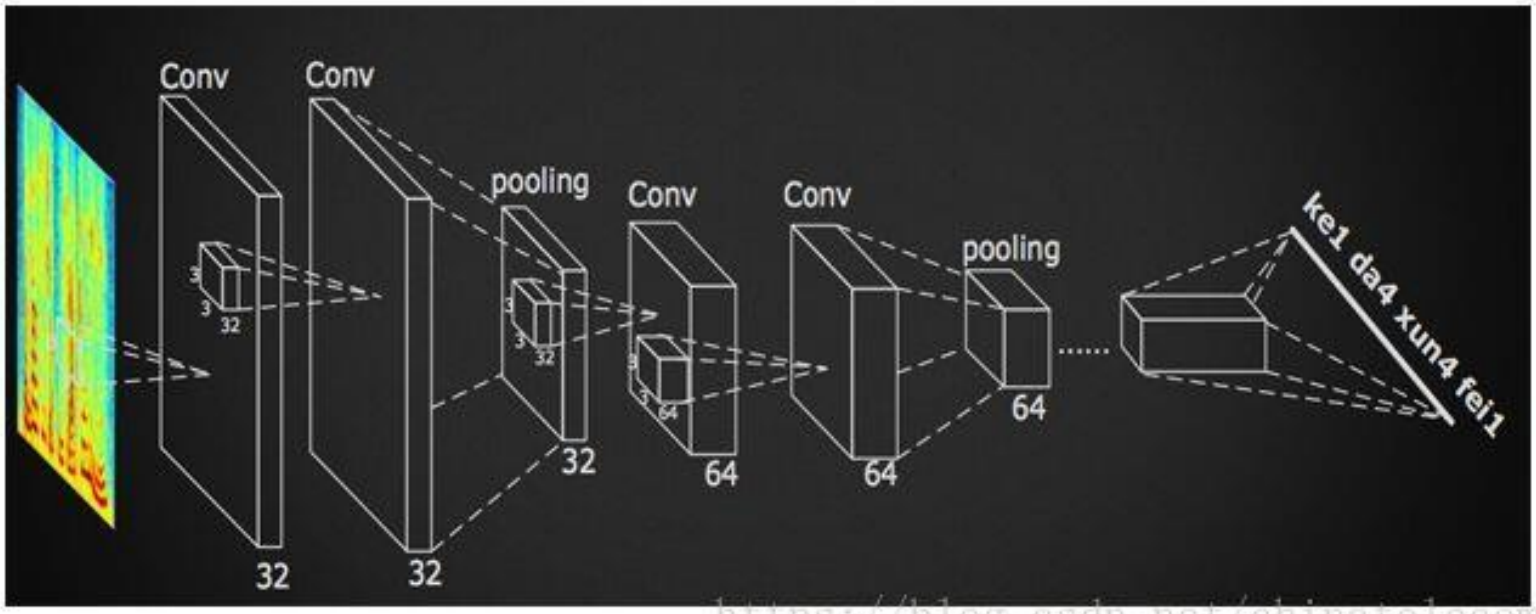
基于Attention的语言模型



- LAS方法
  - 将语音型号的特征输入到双向RNN中(encoder)
  - 在不同时刻关注输入的不同部分（Decode部分）
  - 解码



DFCNN方法



科大讯飞提出的基于深度学习的中文语音识别系统框架，就是DFCNN

开源的语音交互平台

- **CMU-Sphinx**
  - 卡内基 - 梅隆大学开发
- HTK（Hidden Markov Model Toolkit）
  - 剑桥大学工程学院开发
- **Julius**
  - 日本LVCSR研究
- **RWTH ASR**
  - RWTH Aachen 大学
- 其他开源语音识别
  - Kaldi 、simon
  - iATROS-speech 、SHoUT 、Zanzibar OpenIVR等

Toolkit	Programming languages	Development activity	Tutorials and examples	Community	Trained models
CMU Sphinx	Java, C, Python, others	+++	+++	+++	English plus 10 other languages
Kaldi	C++, Python	+++	++	+++	Subset of English
HTK	C, Python	++	+++	++	None
Julius	C, Python	++	++	+	Japanese
ISIP	C++	++	++	+	Digits only

未来的发展趋势

- 语音识别技术
  - 包括特征参数提取技术
  - 模式匹配及模型训练技术
    - 中小词汇量对非特定人语音识别系统识别精度已经大于98%
- 未来语音智能
  - 嘈杂环境麦克风语音、方言识别
  - 移动设备下的语音识别
  - 未来的智能家居系统

练习2：

实现一个简单的声音识别算法，结合提供的声音库文件，实现不同动物的名称识别。

实现语言不限，可以用外部的识别库函数，要求有测试识别率结果分析。

- END