# 语音识别

## 隐马尔科夫模型与序列生成模型

肖睿

四川大学数学学院

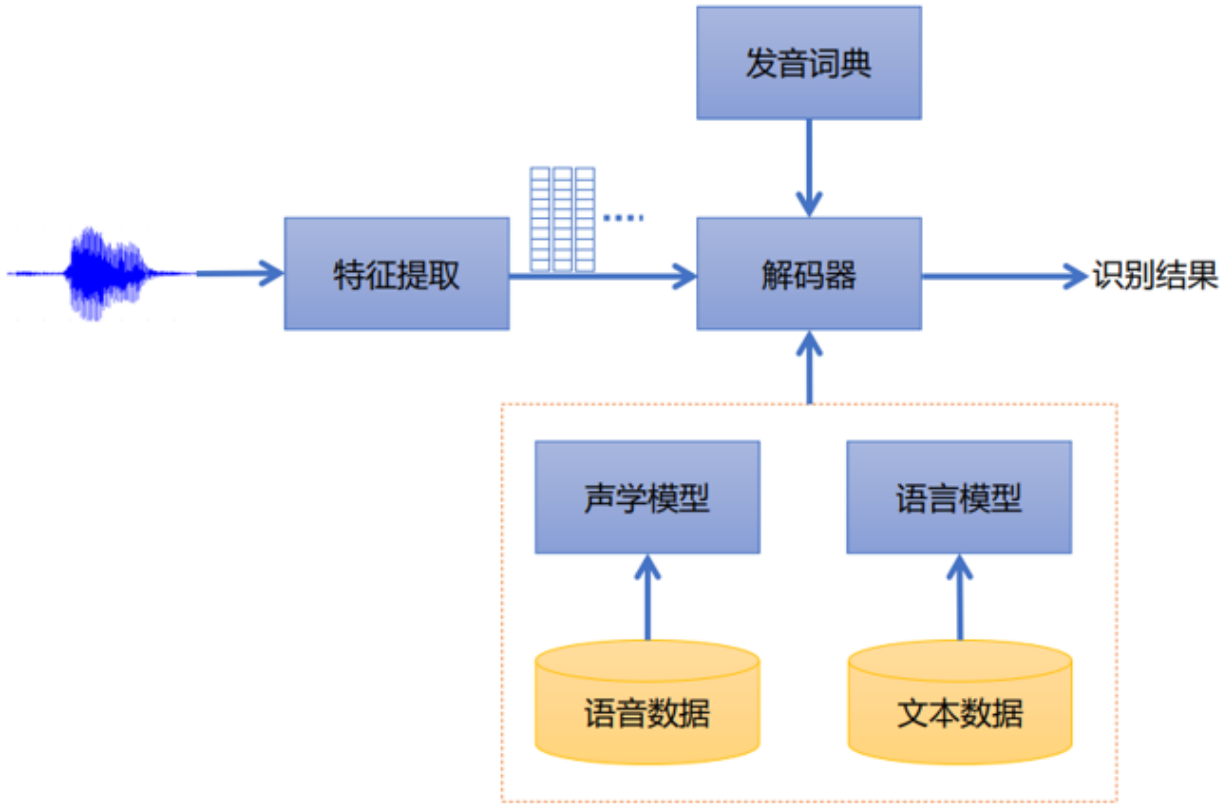
2024 年 6 月 11 日

1. 语音识别
2. 语音合成
3. 语音增强
4. 语音转换
5. 情感语音

1 语音识别
隐马尔科夫模型
识别方法
2 语音合成
3 语音增强
4 语音转换
5 情感语音

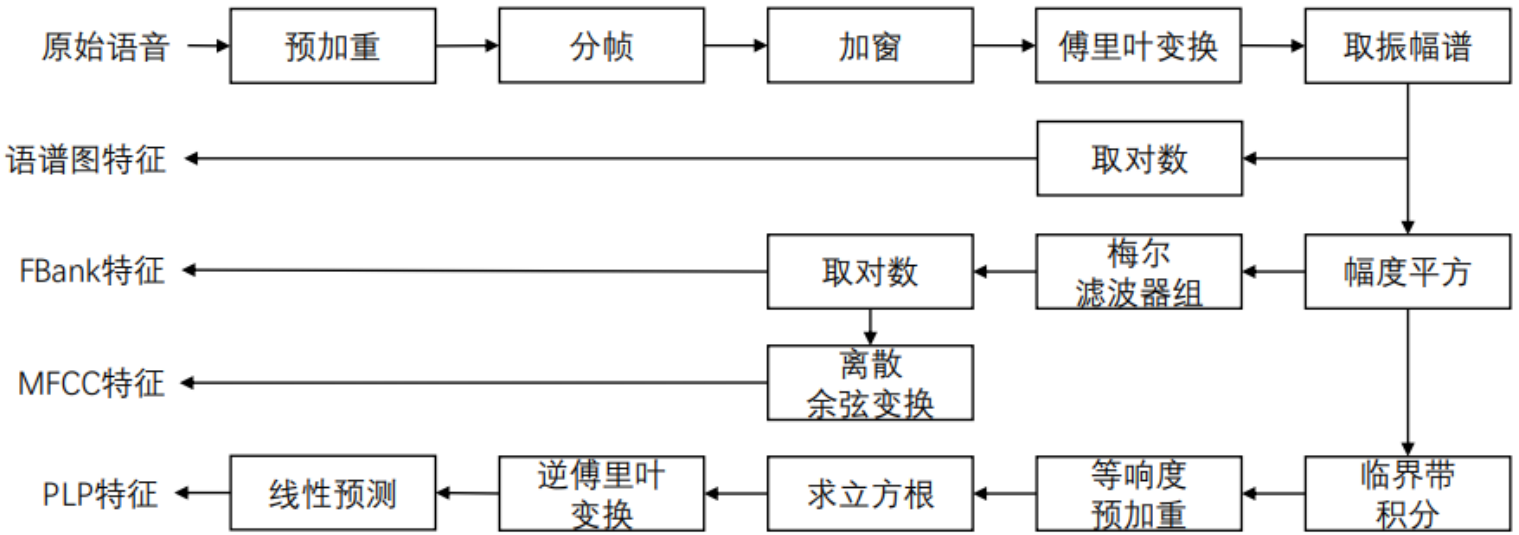
# 系统组成

- 语音识别是指将语音自动转换为文字的过程。语音识别系统主要包括四个部分: 特征提取、声学模型、语言模型和解码搜索。



# 特征提取

- 语音特征抽取即是在原始语音信号中提取出与语音识别最相关的信息，滤除其他无关信息。



# 声学模型

- 1950s　贝尔实验室研制的识别十个英文数字的识别系统
- 1960s　基于模板匹配的方法
- 1980s　基于高斯混合模型-隐马尔科夫模型的技术
- 2010s　基于深度神经网络-隐马尔科夫模型的技术

1 语音识别
隐马尔科夫模型
识别方法
2 语音合成
3 语音增强
4 语音转换
5 情感语音

- 假定系统是一个马尔可夫过程，但其中的状态是“隐藏”的，也就是不直接可见. 我们观察到的是与状态相关的一系列输出或者说“观测值”.
- 自然语言处理的语言识别
- 生物信息学中的 DNA 或蛋白质序列
- 金融领域的股票市场

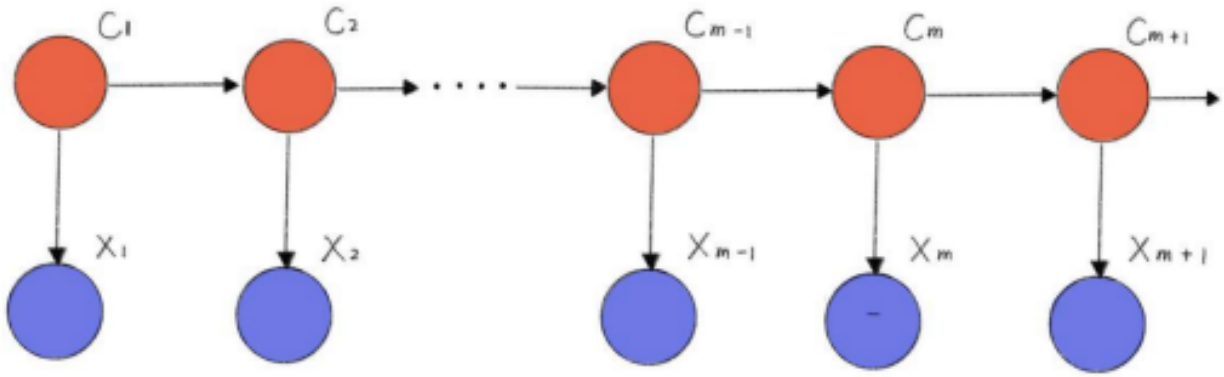



图 1: 马氏链示意图

- 隐马尔科夫链的三个基本问题：
  给定参数，求某一观测序列发生的概率；
  求观测序列发生概率最大时的参数；
  给定观测序列，求最合适的隐藏状态序列。
- 三个基本问题对应隐马尔科夫模型中三个经典的算法：
  向前/向后算法
  Baum-Welch 算法
  Viterbi 算法

# 高斯混合模型-隐马尔科夫模型

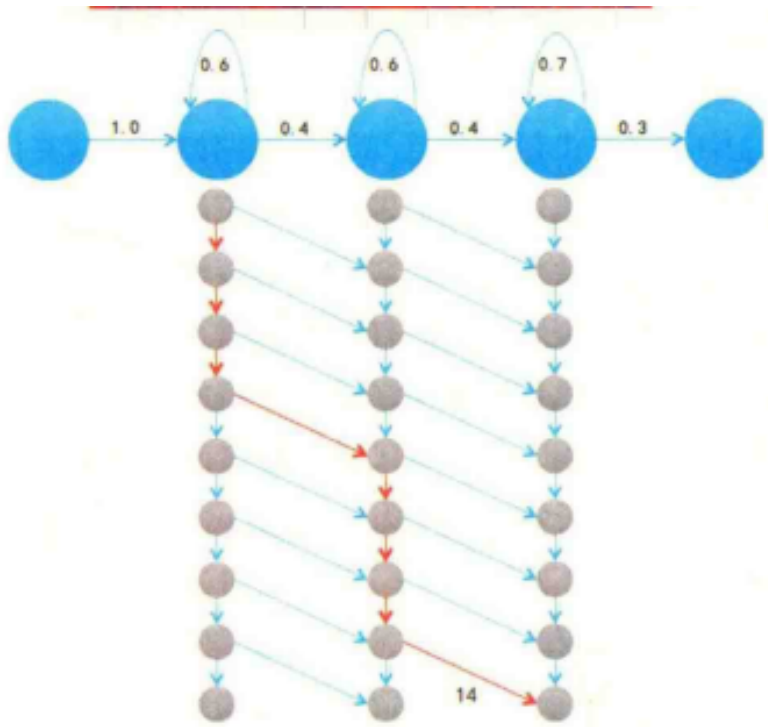



图 2: 隐马尔科夫模型

- 隐马尔科夫模型的参数主要包括状态间转移概率以及每个状态的出现概率。
- 先将每一帧语音带入一个状态计算出现概率，利用不同状态之间的转移概率计算最优路径，该路径对应的概率值为输入语音经隐马尔科夫模型得到的概率值。
- 每个音节训练一个隐马模型，语音只需带入每个音节的模型算一遍，概率最高的音节为识别结果。
- 出现概率采用高斯混合模型

# 高斯混合模型-隐马尔科夫模型

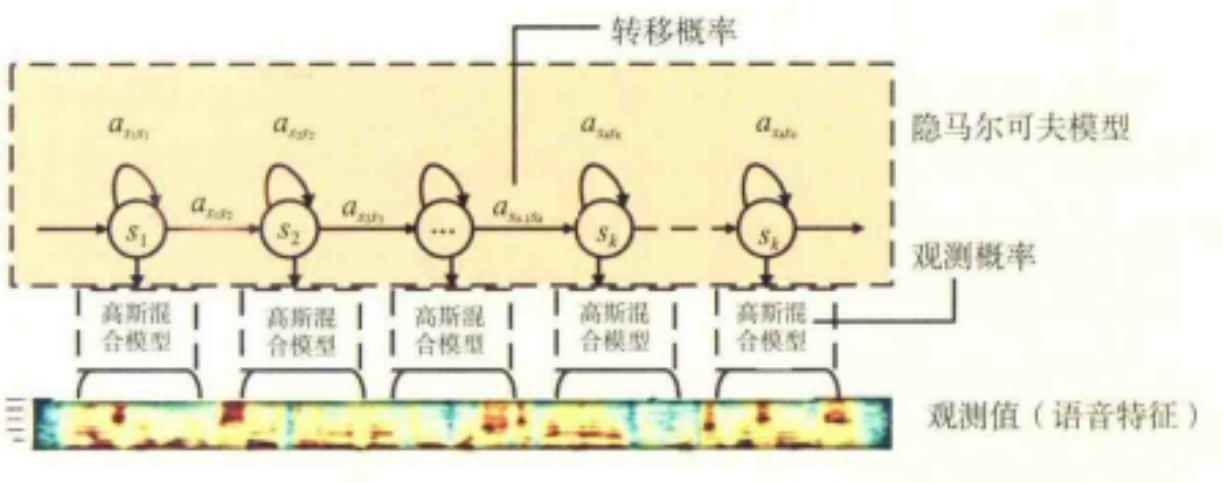



图 3: 高斯混合模型-隐马尔科夫模型的声学模型

- 高斯混合模型不能利用帧的上下文信息，缺乏深层非线性特征变化的内容，如音色差异和发音习惯差异下的相同音。
- 因此只对小词汇量的语音识别任务，使用上下文无关的音素作为建模单元。
- 如图，高斯混合模型估计语音特征的观测概率，隐马尔科夫模型用于描述语音信号的动态变化。

# 向前算法

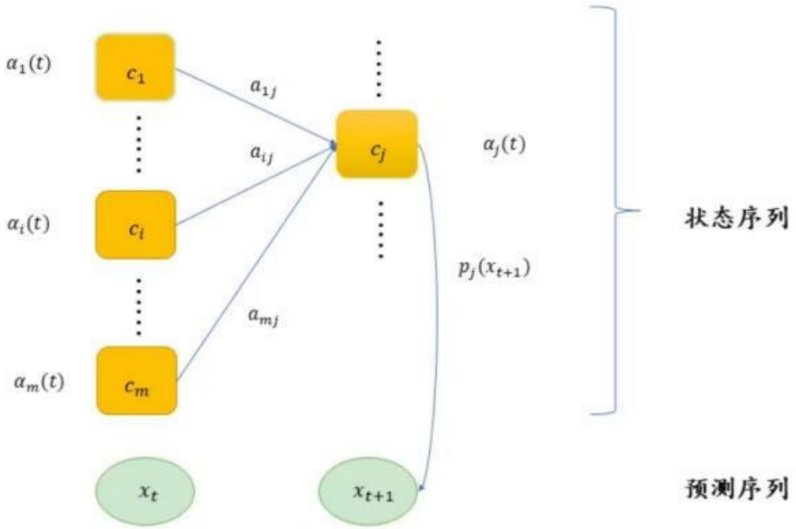



图 4: 向前算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：
   $\alpha_i(1) = \pi_i p_i(x_1), i = 1, 2, ..., m$

# 向前算法

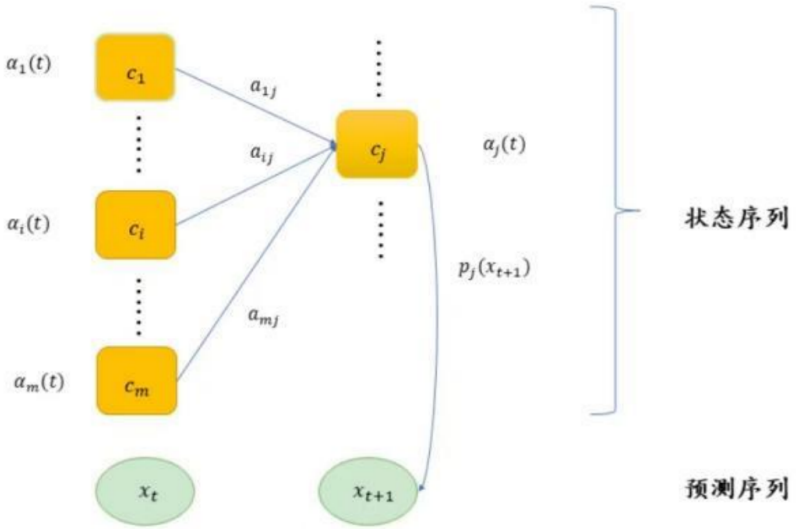



图 4: 向前算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：
   $\alpha_i(1) = \pi_i p_i(x_1), i = 1, 2, ..., m$
2. 递归：$\alpha_j(t+1) = (\sum_{i=1}^m \alpha_i(t) \cdot a_{ij} \cdot p_j(x_{t+1}), t = 1, ..., T-1; i = 1, ..., m$

# 向前算法

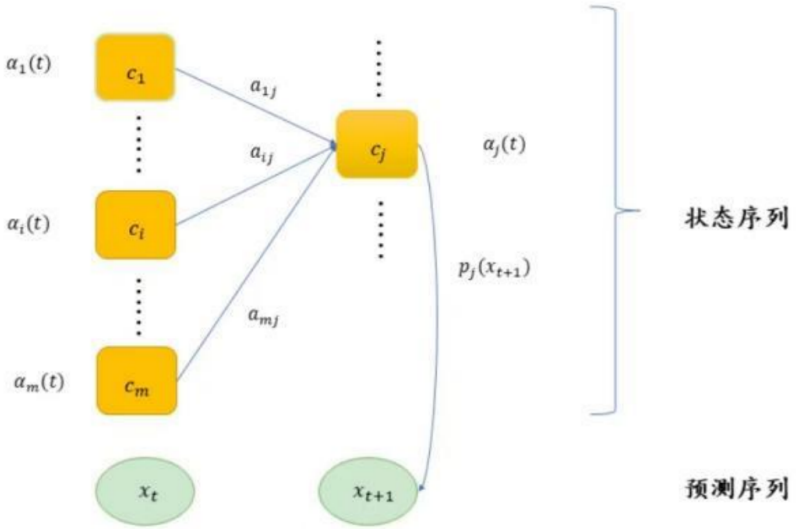



图 4: 向前算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：
   $\alpha_i(1) = \pi_i p_i(x_1), i = 1, 2, ..., m$
2. 递归：$\alpha_j(t+1) = (\Sigma_{i=1}^m \alpha_i(t) \cdot a_{ij} \cdot p_j(x_{t+1}), t = 1, ..., T-1; i = 1, ..., m$
3. 代入终值计算：
   $L = P(X_1^T = x_1^T) = \Sigma_{i=1}^m \alpha_i(T)$

# 向后算法

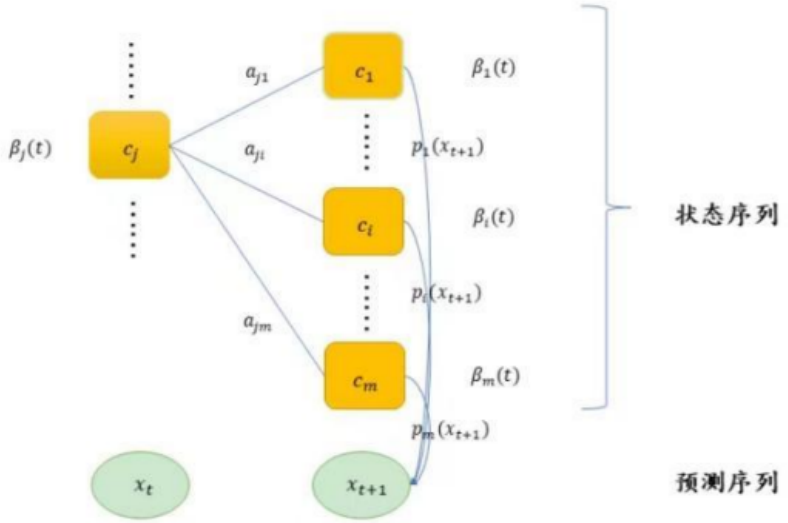



图 5: 向后算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：$\beta_i(T) = 1, i = 1, 2, ..., m$

# 向后算法

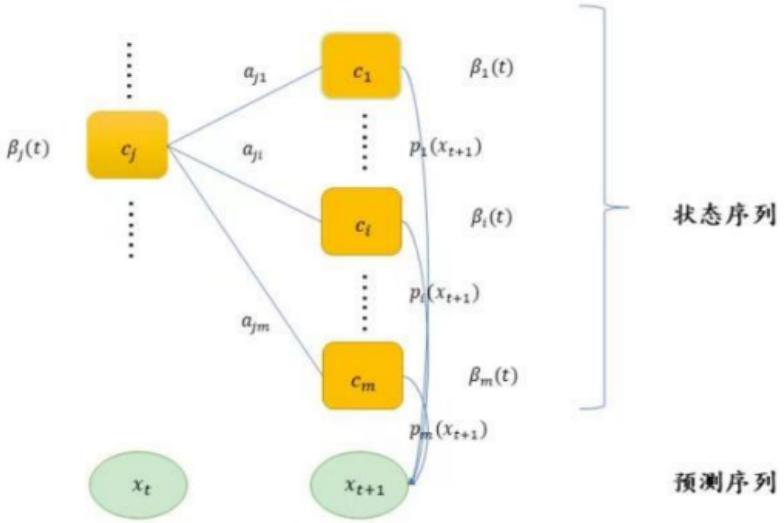



图 5: 向后算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：$\beta_i(T) = 1, i = 1, 2, ..., m$
2. 递归：$\beta_j(t) = (\Sigma_{i=1}^m \beta_i(t+1) \cdot a_{ji} \cdot p_j(x_{t+1}), t = 1, ..., T-1; i = 1, ..., m$

# 向后算法

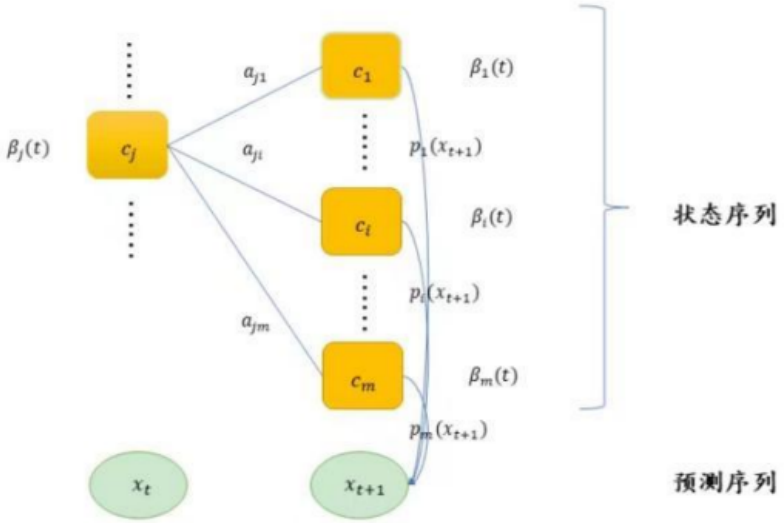



图 5: 向后算法原理图

给定参数 $\theta$，计算概率 $P(X_1^t = x_1^t|\theta)$ 的步骤：

1. 初始化：$\beta_i(T) = 1, i = 1, 2, ..., m$
2. 递归：$\beta_j(t) = (\Sigma_{i=1}^m \beta_i(t+1) \cdot a_{ji} \cdot p_j(x_{t+1}), t = 1, ..., T-1; i = 1, ..., m$
3. 代入终值计算：$L = P(X_1^T = x_1^T) = \Sigma_{i=1}^m \pi_i \cdot p_i(t) \cdot \beta_i(1)$

# Baum-Welch 算法

- 本算法期望最大化算法在隐马尔科夫模型中的特殊形式，是极大似然估计的延伸，用于求解含有不可观测变量的参数估计问题。
- 对数似然函数的条件期望 $Q(\theta, \tilde{\theta}) = \mathbf{E}[\log p(x_1^T; y_1^T|\theta))|x_1^T, \theta] = \Sigma_{y_1,\ldots,y_T} \log p(x_1^T; y_1^T|\theta) \cdot p(y_1^T|x_1^T; \tilde{\theta})$，最大化 $Q$ 得到 $\theta$ 的估计值。

# Viterbi 算法

- 要寻找一个隐藏状态序列，能最佳地解释给定的观测序列，数学地来说，要极大化 $P(C_1^T = c_1^T | X_1^T = x_1^T)$。Viterbi 算法可以找到这样的序列。
- 定义 $V_i(t) = max_{c_1,...,c_{t-1}} P(C_1^{t-1} = c_1^{t-1}, C_t = i, X_1^t = x_1^t) \quad (t = 2, ..., T) \quad V_i(1) = \pi_i \cdot p_i(x_1)$ 是初值条件。
- 先引入 $V_j(t)$: $V_j(t) = \left(max_i V_i(t-1)_{ij}\right) \cdot p_j(x_t) \quad (t = 2, ..., T, i = 1, ...m)$
- 利用维特比公式计算最优序列:
  $s_T = argmax_i V_i(T) \quad s_t = argmax_i(V_i(t) \cdot a_{i,s_{t+1}}) \quad (t = T-1, ..., 1)$

# 算法演示：利用 HMM 合成古典音乐

- 数据集格式音符序列：

'C4', 'C3', 'D#4', 'D#3', 'G#4', 'G#3', 'A#3', 'A#4', 'C5', 'C4', 'G#4', 'G#3', 'D#4', 'D#5', 'A#4', 'A#5', 'C6', 'C5', 'G#5', 'G#4', 'D#6', 'D#5', 'A#5', 'A#6', 'C6', 'C7', 'G5', 'G6', 'A#5', 'A#6', 'G#5', 'G#6', 'G5', 'G6', 'F#6', 'F#5', 'F#5', 'F#6', 'G5', 'G6', 'F#6', 'F#5', 'F5', 'F6', 'F#5', 'F#6', 'A5', 'A6', 'G5', 'G6', 'D#5', 'D#6', 'D#5', 'D#6', 'D5', 'D6', 'F5', 'F6', 'D#6', 'D#5', 'D5', 'D6', 'D5', 'D6', 'G4', 'C5', 'C6', 'D#4', 'G5', 'D4', 'G4', 'D#5', 'A#5', 'F#5', ……

- 隐藏状态为作曲家的创作动机；观测状态为作曲家写下的音符。
- 用音符序列作为观测值，通过极大似然估计得到状态转移矩阵、发射矩阵、初始分布的估计。

# 深度神经网络-隐马尔科夫模型

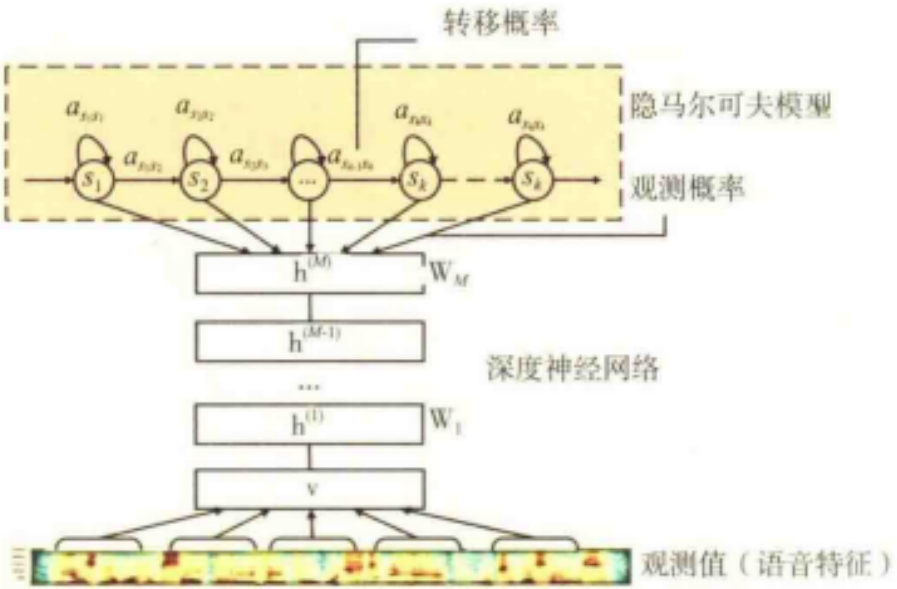



图 6: 深度神经网络-隐马尔科夫模型的声学模型

- 建模单元为聚类后的三音素状态。
- 建模单元可以是音素、音节、词语等。词语长度不等、粒度大，然而词语包含的音素是确定且有限的，故以音素建模。
- 而音素与上下文相关，故一般使用三音素进行建模。
- 三因素数量庞大，用决策树对三音素聚类以减少数目。
- 本模型能利用上下文信息且能学习非线性的更高层次特征表达。

# 语言模型

- N 元文法语言模型和循环神经网络语言模型。
- 评价指标是语言模型在测试集上的困惑度，即句子不确定性的程度。困惑度越小指我们对该句子的理解的程度越深，越接近真实语言的分布。

# 解码搜索

- 解码搜索时在由声学模型、发音词典、语言模型构成的搜索空间中寻找最佳路径。
- 解码需要声学得分与语言得分。每一帧都有声学得分，但词语级别才有语言得分，一个词覆盖多帧语言特征。引入参数平滑语言得分，使两种得分有相同尺度。
- 构建解码空间的方法有静态和动态两类。
- 搜索算法分两类：时间同步方法，如维特比算法；时间异步方法，如 A 星算法。

1 语音识别
隐马尔科夫模型
识别方法
2 语音合成
3 语音增强
4 语音转换
5 情感语音

# 基于端到端的语音识别方法

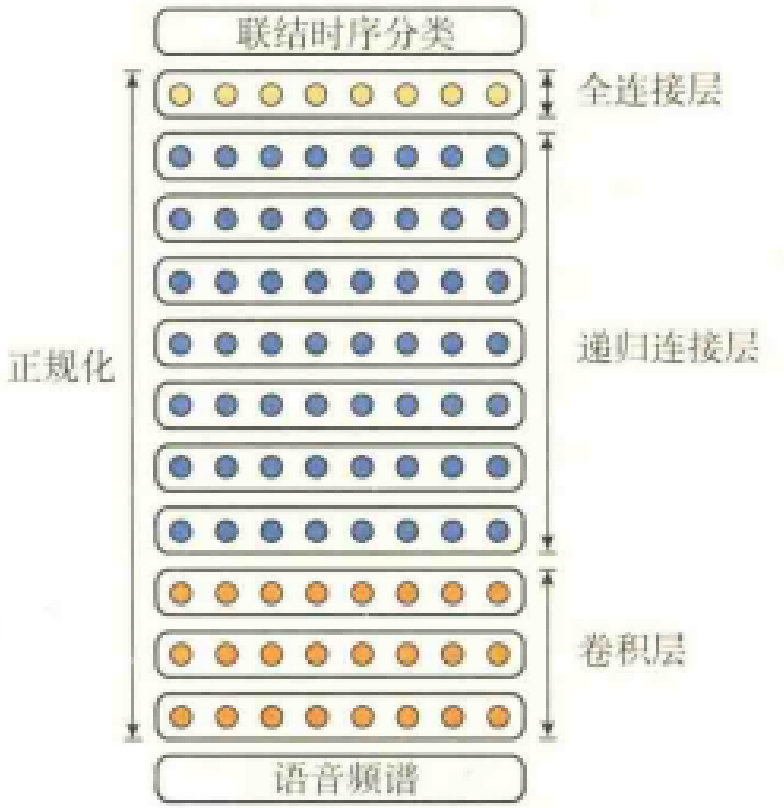



图 7: 基于联结时序分类的端到端声学模型

- 为了解决训练繁复、性能受限于模型精度等问题，提出了端到端的识别方法。第一类是基于联结时序分类的端到端的声学建模方法。
- 引入训练准则联结时序分类，损失函数的优化目标是输入和输出在句子级别对齐，而不是帧级别对齐，只需对从输入到输出的序列映射关系建模。
- 循环神经网络用来序列建模，联结时序分类损失函数与长短时记忆结合使用，或卷积神经网络。
- 建模单元可以是音素或者是字，粒度改变增加了语音数据的冗余度，提高了音素区分度，也增加了解码速度。

# 基于端到端的语音识别方法

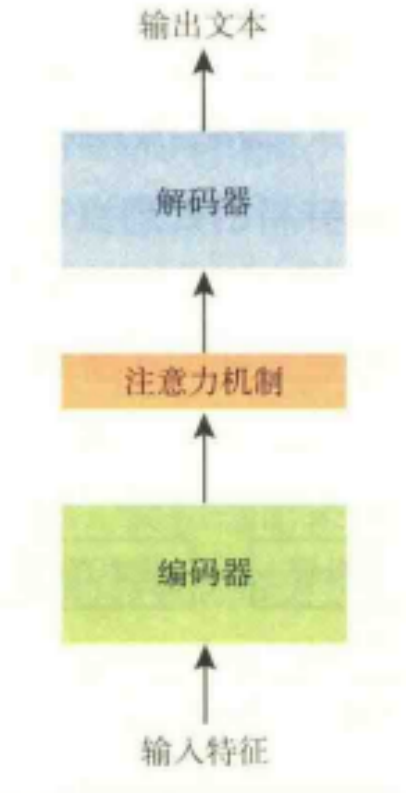



图 8: 基于注意力机制的端到端识别系统结构

- 第二类是基于注意力机制的端到端语音识别系统。
- 将声学模型、发音词典和语言模型联合为一个模型进行训练。
- 编码器将不定长的输入序列映射成定长的特征序列，注意力机制提取序列中信息，解码器将定长序列扩展为输出的单元序列。

1 语音识别

2 语音合成
序列生成模型

3 语音增强

4 语音转换

5 情感语音

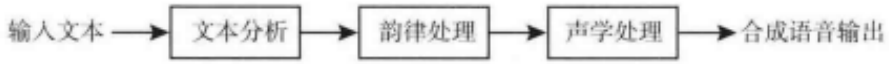



图 9: 语音合成系统

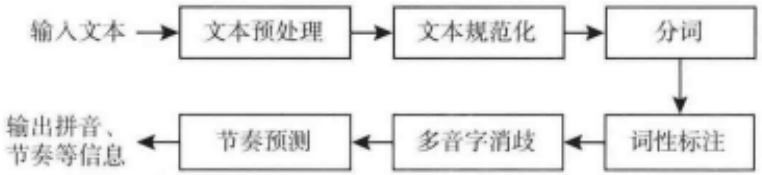



图 10: 文本分析流程

- 文本分析模块尽可能多地输出语言学信息（拼音、节奏等），实际上是一个人工智能系统，属于自然语言理解的范畴。

# 基于拼接的语音合成

- 根据文本分析的结果，从预录制并标注好的语音库中挑选单元进行调整并拼接。基本元是音节或音素。
- 大语料库具有较高的上下文覆盖率，使基本元几乎不作调整就可以拼接。

# 基于参数的语音合成

- 基于隐马尔科夫模型的语音合成方法分为训练和合成两个阶段。
- 训练前，配置建模参数：单元大小、状态数目、模型拓扑结构等。
- 训练数据包括语音数据和标注数据（包括音段切分和韵律标注）
- 设计上下文属性集和用于决策树聚类的问题集，如对后调、前后生韵母等影响上下文属性的的语音参数设计问题集。
- 决策树聚类模型精度低，可以将隐马尔科夫模型替换为深度学习网络来预测声学参数。

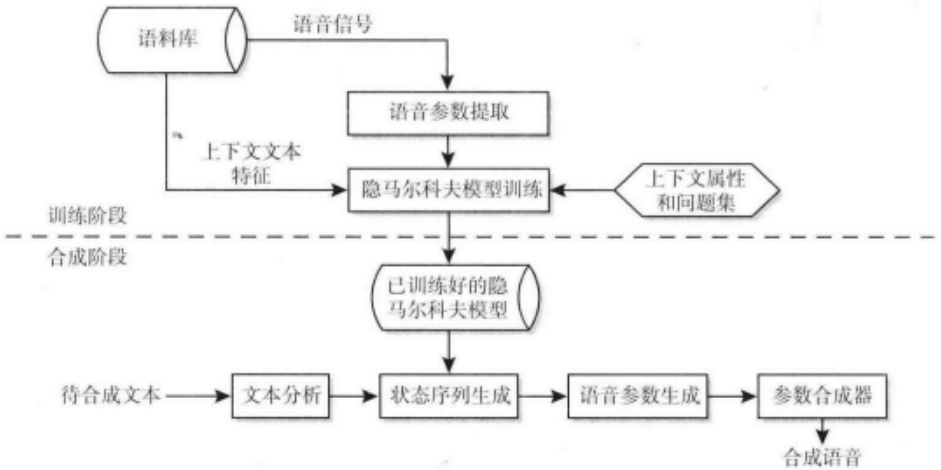



图 11: 隐马尔科夫模型的语音合成系统框架图

# 基于端到端的语音合成

- 从字符或者音素直接合成语音，克服了参数化多段建模的缺陷。
- 编码器是以字符或音素为输入的神经网络；解码器是带有注意力机制的循环神经网络，输出文本序列或音素序列的频谱图。

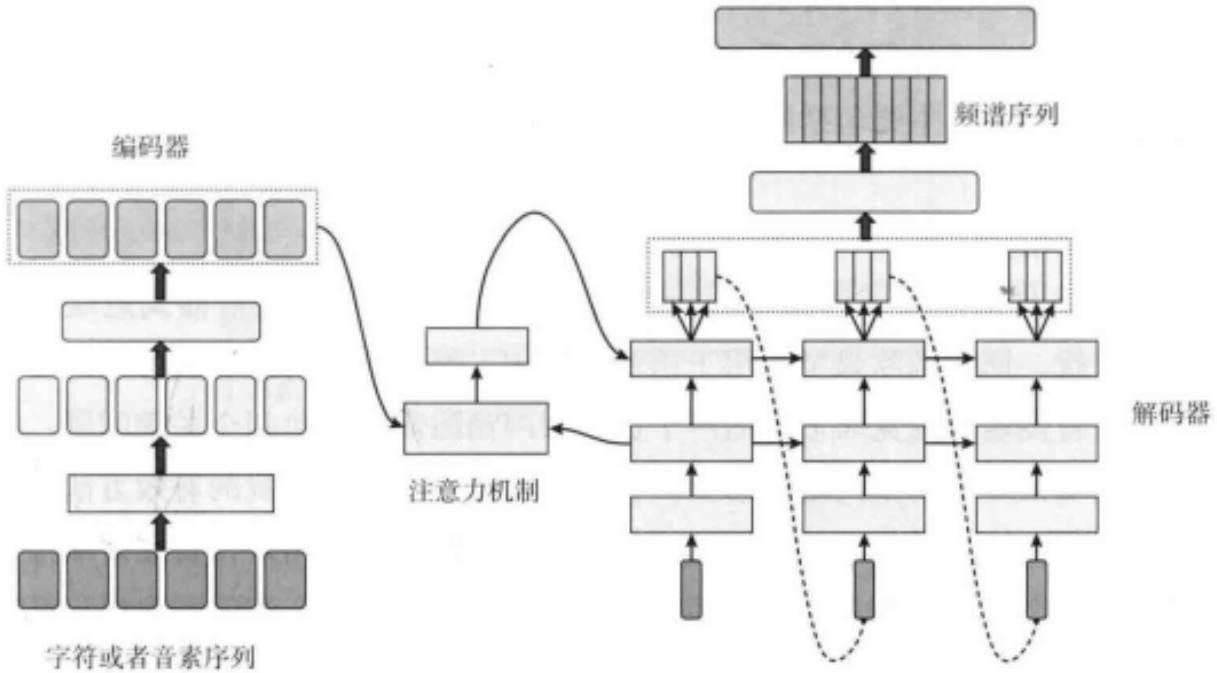



图 12: 端到端语音合成框架图

1 语音识别

2 语音合成
序列生成模型

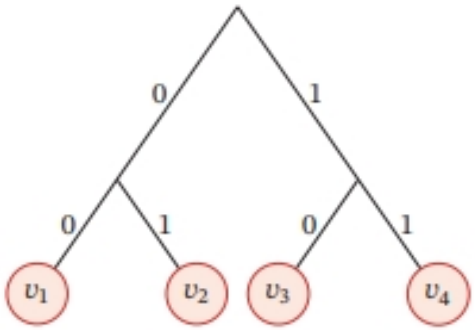
3 语音增强

4 语音转换

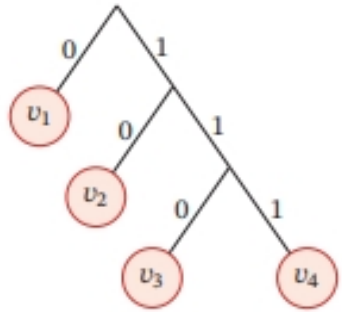
5 情感语音

# 聚类问题集

- 层次化 Softmax：将单元分为 $K$ 组，提高计算效率。
- 用树结构来组织问题集：二叉树或霍夫曼编码，将每个单元编码为二值变量的位向量。




(a) 平衡树




(b) 霍夫曼编码树

# Seq2Seq 模型

序列到序列模型的目标是估计条件概率

$$p_\theta(\mathbf{y}_{1:T}|\mathbf{x}_{1:S}) = \prod_{t=1}^{T} p_\theta(y_t|\mathbf{y}_{1:(t-1)}, \mathbf{x}_{1:S}),$$

其中 $y_t \in \nu$ 为词表 $\nu$ 中的某个词. 给定一组训练数据 $\{(x_{S_n}, y_{T_n})\}_{n=1}^{N}$, 我们可以使用最大似然估计来训练模型参数

$$\hat{\theta} = \arg\max \sum_{n=1}^{N} \log p_\theta(\mathbf{y}_{1:T_n}|\mathbf{x}_{1:S_n}).$$

一旦训练完成, 模型就可以根据一个输入序列 $x$ 来生成最可能的目标序列.

$$\hat{\mathbf{y}} = \arg\max_{y} p_{\hat{\theta}}(\mathbf{y}|\mathbf{x}),$$

# 基于循环神经网络的 Seq2Seq 模型

编码器：首先使用一个循环神经网络 $f_{\text{enc}}$ 来编码输入序列 $x_{1:S}$ 得到一个固定维度的向量 $u$，通常为编码循环神经网络最后时刻的隐状态：

$$\boldsymbol{h}_t^{\text{enc}} = f_{\text{enc}}(\boldsymbol{h}_{t-1}^{\text{enc}}, \boldsymbol{e}_{x_{t-1}}, \theta_{\text{enc}}), \quad \forall t \in [1:S], \tag{1}$$

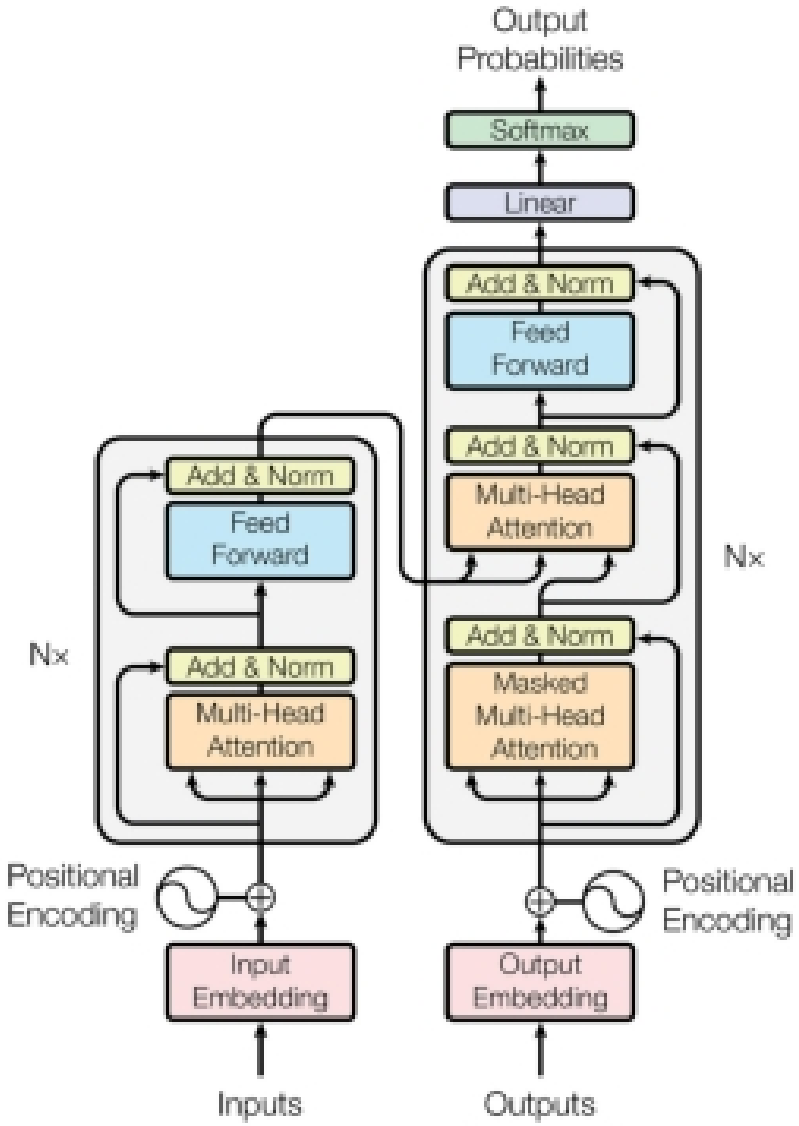
$$\boldsymbol{u} = \boldsymbol{h}_S^{\text{enc}}, \tag{2}$$

解码器：在生成目标序列时，使用另外一个循环神经网络 $f_{\text{dec}}$ 来进行解码。在解码过程的第 $t$ 步时，已生成前缀序列为 $y_{1:(t-1)}$。令 $\boldsymbol{h}_t^{\text{dec}}$ 表示在网络 $f_{\text{dec}}$ 的隐状态，则：

$$\boldsymbol{h}_0^{\text{dec}} = \boldsymbol{u}, \tag{3}$$

$$\boldsymbol{h}_t^{\text{dec}} = f_{\text{dec}}(\boldsymbol{h}_{t-1}^{\text{dec}}, \boldsymbol{e}_{y_{t-1}}, \theta_{\text{dec}}). \tag{4}$$

# Transformer 模型



1 语音识别

2 语音合成

3 语音增强

4 语音转换

5 情感语音

- 语音增强：当语音信号被各种给与干扰源研磨后，从混叠信号中提取出有用的语音信号，抑制、降低各种干扰。
- 主要包括回声消除、混响抑制、语音降噪等关键技术。

# 回声消除

- 回声干扰：远端扬声器播放的声音传播到近端的麦克风形成的干扰。
- 回声消除的关键：
  (1) 远端信号和近端信号的同步问题；
  (2) 双讲模式下消除回波信号干扰的有效方法。
- 应用中面临的问题：回声消除算法将扬声器作为信号源，但扬声器播放时的非线性失真、传播过程中的衰减、噪声干扰、回声干扰等问题。

# 混响抑制

- 声音在房间传输时会被墙壁和障碍物反射，再通过不同路径对麦克风形成干扰源。
- 声源停止发声后，声压级减少 60 分贝所需要的时间为 T60 混响时间。
- 去混响时，更多关注于抑制晚期混响。

# 语音降噪

- 噪声抑制可以分为单通道的语音降噪和多通道的语音降噪，前者通过单个麦克风去噪，后者通过麦克风列阵算法增强目标方向的声音。
- 多通道语音降噪的核心问题是估计空间滤波器，即输入麦克风阵列采集的多通道信号，输出处理后的单路语音信号。
  麦克风阵列结构影响算法效果。经典的麦克风阵列包括线阵（多用于智能车载系统）和环阵（智能音箱系统）；
  麦克风数量增多加强噪声抑制能力但影响算法精度，因此广泛采用双麦结构。
- 单通道语音降噪方法主要包括：
  (1) 基于信号处理技术的方法（处理平稳噪声）；
  (2) 基于矩阵分解的方法（计算复杂度高）；
  (3) 基于数据驱动的方法（性能依赖训练集与测试集的匹配度）；
  (4) 基于深度学习的语音降噪方法（处理非平稳噪声、鲁棒性更高）。

1 语音识别

2 语音合成

3 语音增强

4 语音转换

5 情感语音

- 语音转换首先提取说话人身份相关的声学特征参数，然后用改变后的声学特征参数合成出目标人说话的语音。
- 训练阶段，首先提取源说话人和目标说话人的个性特征参数，根据某种匹配规则简历二者之间的匹配函数；
- 转换阶段，利用匹配函数转换特征参数，再用新的特征参数合成语音。

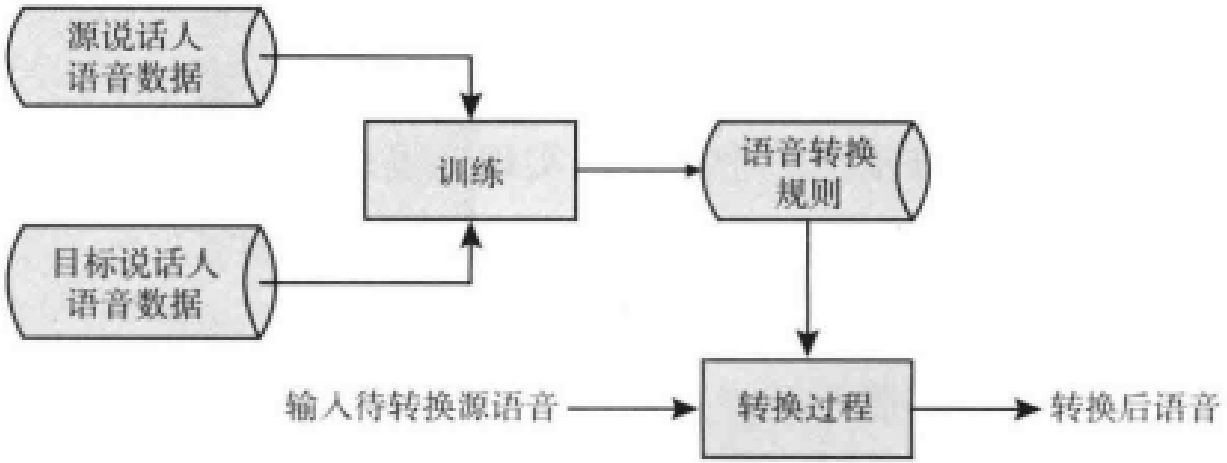



图 13: 语音转换系统框架图

- 码本映射法：源码本与目标码本的单元一一对应，码本是语音片段中的关键语音帧，但转换函数不连续。
- 高斯混合模型法：用统计参数模型建立映射关系，用最小均方误差准则来确定转换函数，转换函数的参数由参数估计算法得到，但转换特征过平滑。
- 深度神经网络法：非线性建模建立映射关系，实现个性信息的转化，解决过平滑问题，还具有能处理高维数据、模型能快速自适应、隐私保护等优点。

1 语音识别

2 语音合成

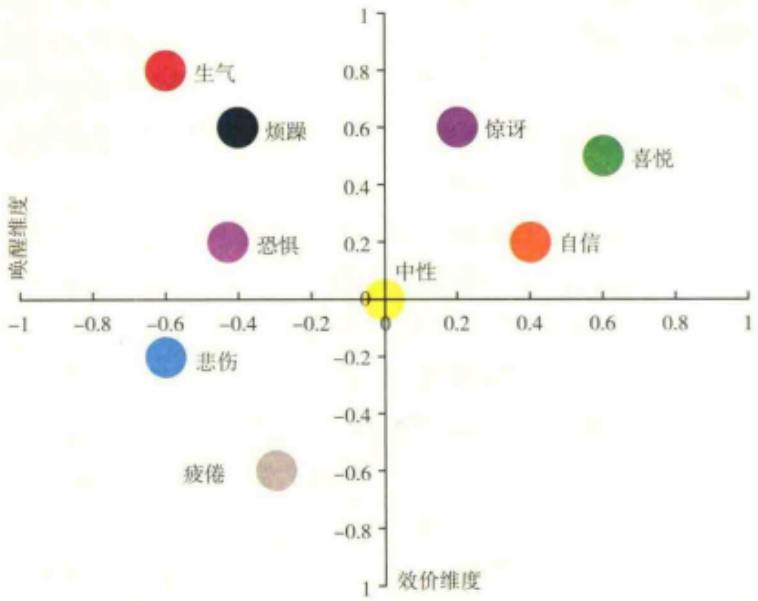
3 语音增强

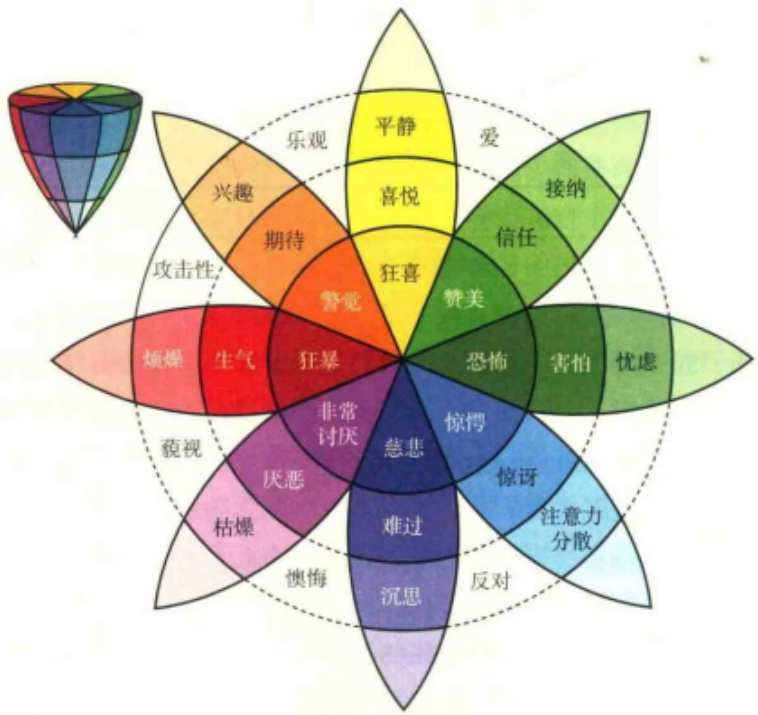
4 语音转换

5 情感语音

# 情感描述

- 离散情感：由形容词标签形式描述，如六大基本情感。
- 维度情感：由情感空间描述。情感空间是笛卡尔空间，每一维对应一个心理学属性（激活度属性、愉悦度属性等）。



# 声学特征

- 语音声学情感特征分为三类：韵律特征、音质特征、频谱特征。
- 这三类特征不同语段长度的统计特征也被普遍使用，如均值、变化率、变化范围等。

| | 愤怒 | 高兴 | 悲伤 | 恐惧 | 厌恶 |
|---|---|---|---|---|---|
| 语速 | 略快 | 快或慢 | 略慢 | 很快 | 非常快 |
| 平均基音 | 非常高 | 很高 | 略低 | 非常高 | 非常低 |
| 基音范围 | 很宽 | 很宽 | 略窄 | 很宽 | 略宽 |
| 强度 | 高 | 高 | 低 | 正常 | 低 |
| 声音质量 | 有呼吸声、胸腔声 | 有呼吸声、共鸣音调 | 有共鸣声 | 不规则声音 | 嘟囔声、胸腔声 |
| 基音变化 | 重音处突变 | 光滑、向上弯曲 | 向下弯曲 | 正常 | 宽，最终向下弯曲 |
| 清晰度 | 含糊 | 正常 | 含糊 | 精确 | 正常 |

# 情感识别

- 情感识别系统由语音信号采集、语音情感特征提取、语音情感识别组成。
- 情感识别是一个模式分类问题，HMM、SVM、高斯混合模型等经典算法皆可以使用于情感识别。