



语音信号处理

语音合成基本原理

李军锋

中国科学院声学研究所



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

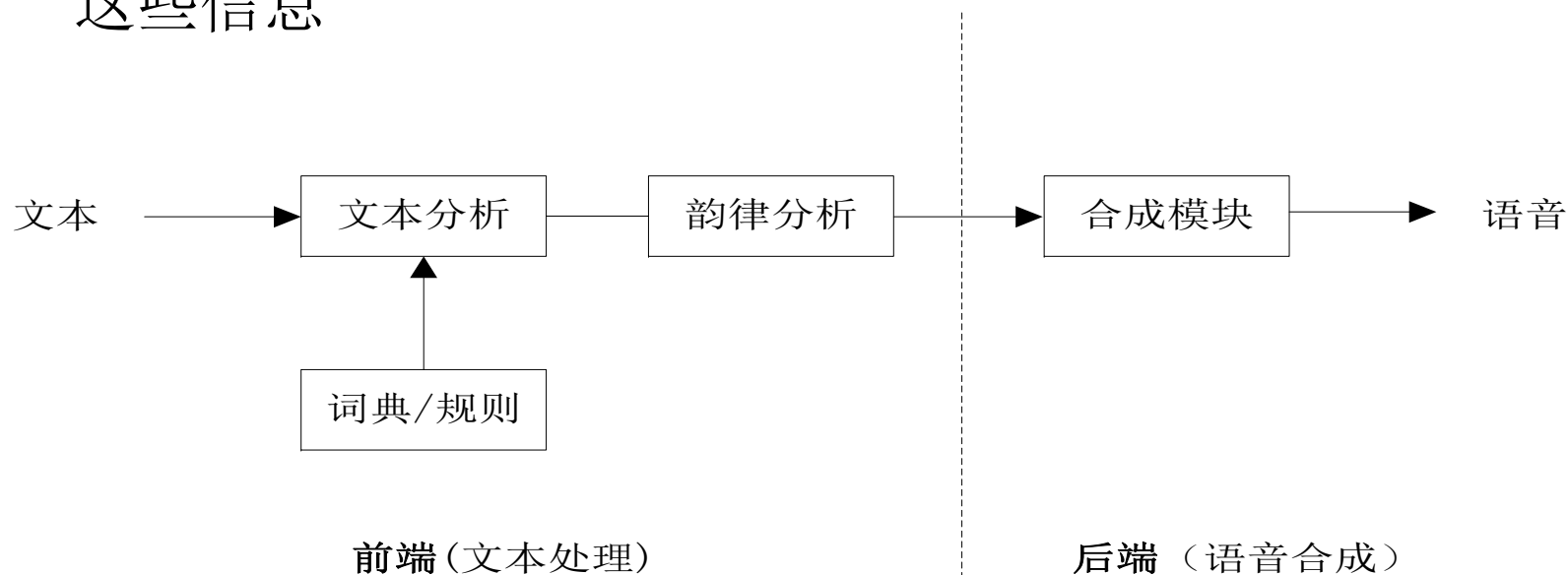
6、语音合成效果评测



语音合成系统概述

什么叫语音合成：

- 语音合成是通过机械的、电子的方法产生人造语音的技术。它的目的是使一些以其他方式表示或存储的信息能转换为清晰易懂的语音，从而让人们能够利用听觉获取这些信息



语音合成系统示意图



语音合成系统概述



语音合成系统示意图



语音合成系统概述

三种主流语音合成系统对比

| 系统 | 优点 | 缺点 |
|------|-----------------------------|------------------------------|
| 波形拼接 | 音质好，自然度高 | 对数据要求高，系统构建复杂，磁盘占用高，合成语音不连贯 |
| 统计参数 | 流畅度高，易懂度高，灵活度高，系统尺寸小，构建系统迅速 | 自然度不够，过于平滑 |
| 端到端 | 自然度高，流畅度高 | 技术不成熟，训练过程复杂，对训练数据要求高，合成速度较慢 |



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



语音合成系统前端

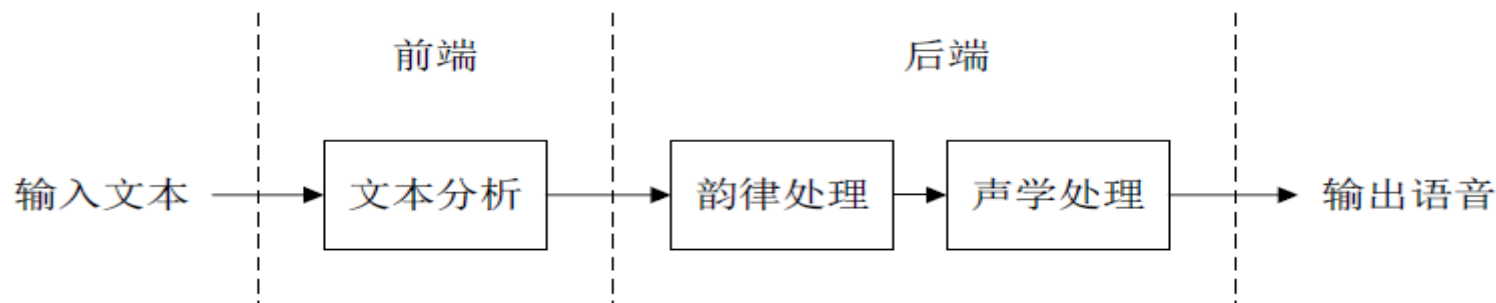


图 1-1 典型的语音合成系统示意图

三个主要模块：

- 文本分析： 文本规范化、分词、字音转换
- 韵律处理： 基频、时长、节奏预测
- 声学处理： 统计参数建模和声码器



语音合成系统前端

文本规范化

- 针对各种数字进行处理
- 针对各种特殊符号进行处理
- 采用基于规则的方法进行文本正则化处理

Mr, Dr, Rd

£5, \$5 million, 12° C

1995 2001 1,995 ☎ 236 3017 233 4488



语音合成系统前端

字音转换模型（多音字问题）

- 采用统计机器学习方法进行字音转换
- 融合词的搭配、词性、句法结构等多种特征

例：

- 银行/行人
- 教授（jiào shòu） / 教授（jiāo shòu）



语音合成系统前端

韵律

- 基频、时长、停顿、重音、能量等

韵律节奏的层级结构

- 韵律词
- 韵律短语
- 语调短语

韵律预测主要上下文特征

- 词性、词频、词长、当前词位置 、句子长度

在完成文本正则化、分词以及特征提取后，使用HMM对韵律（时长）进行建模。



语音合成系统前端

韵律层级划分示例

语调短语：使用程序节省了时间且提高了数据的准确性。

韵律短语：使用程序节省了时间 且提高了数据的准确性。

语调词：使用 程序 节省了 时间 且 提高了 数据的 准确性。

音节：使用 程 序 节 省 了 时 间 且 提 高 了 数 据 的 准 确 性。



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



拼接语音合成

拼接语音合成
关键技术

语料库构建

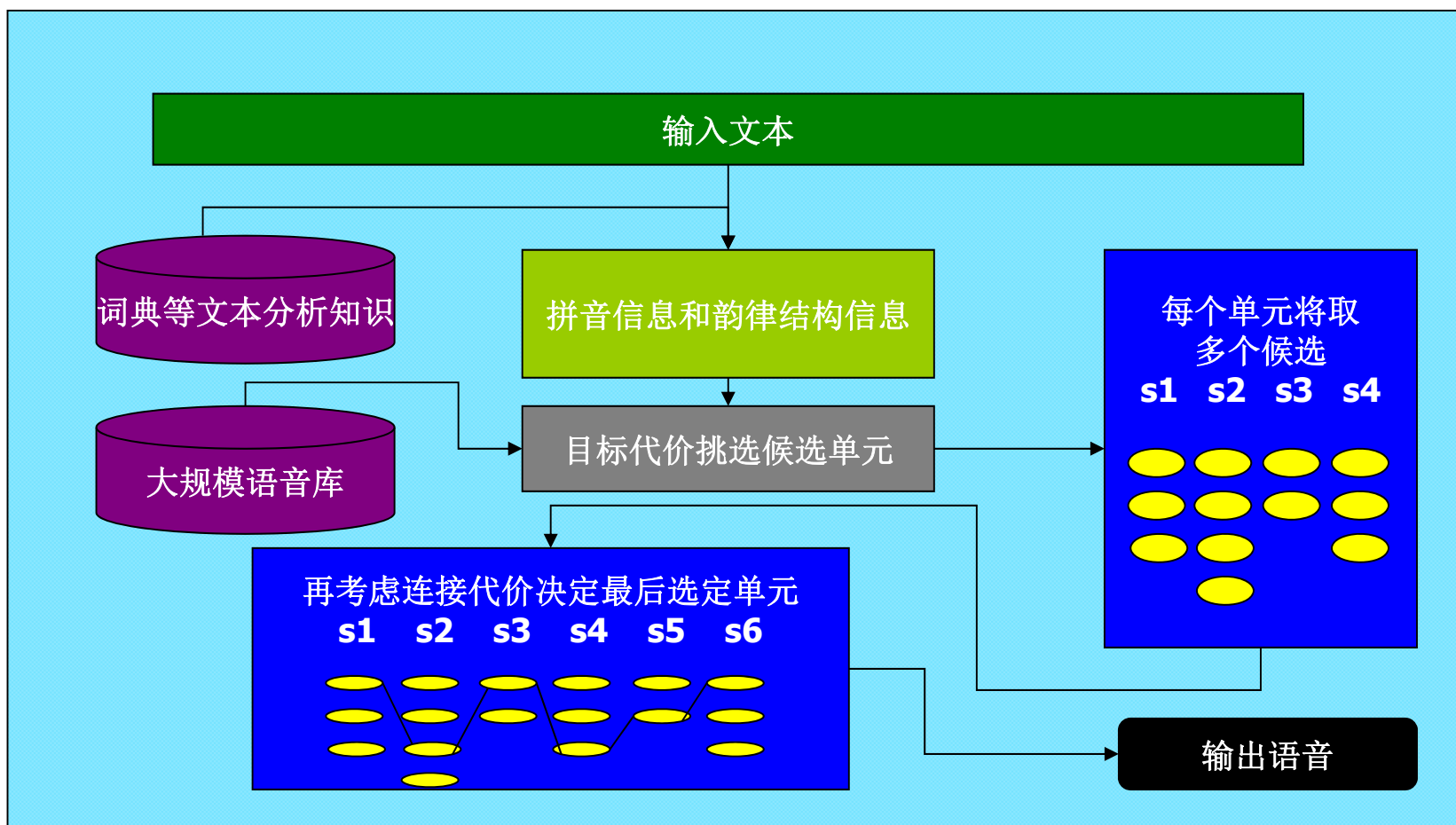
单元选择

波形平滑



拼接语音合成

- 拼接语音合成处理流程





1:语音库构建

由于语境不同，重音表现不同，每个单元的声学特征有很大不同。

一个理想的语音库应该能够覆盖所有的音节特征（左音调、右音调）、韵律（音素位置）特征。



音库规模扩大，则需要更高的成本



2:单元选择

选音过程中往往采用多种复杂的技术，包括多项统计学上的技术或神经网络技

一种衡量单元选择误差的方法：上下文矢量距离
每个单元的音段和韵律上下文形成一个六维向量，
用两个矢量之间的距离表示两个单元的匹配程度



3:波形平滑

核心思想：直接对存储于音库的语音运用PSOLA算法进行拼接从而合成完整的语音

主要特点：有别于传统概念中只是将不同的语音单元进行简单拼接，PSOLA系统首先要在大量语音库中，选择最合适的语音单元用于拼接，使合成波形既保持了原始发音的主要音段特征，又能使拼接单元的韵律特征符合上下文的要求，从而获得很高的清晰度和自然度



3:波形平滑

本质上说，PSOLA算法是利用短时傅里叶变换重构信号的叠加法

信号 $x(n)$ 的短时傅里叶变换为

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)\omega(n-m)e^{-j\omega n}, n \in \mathbf{Z}$$



3:波形平滑

由于语音信号短时平稳，因此在时域每隔若干个（例如R个）样本取一个频谱函数就能重构信号 $x(n)$ ，即可令

$$Y_r(e^{j\omega}) = X_n(e^{j\omega})|_{n=rR}, r, n \in \mathbf{Z}$$



3:波形平滑

上式的傅里叶逆变换为

$$y_r(m) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_r(e^{j\omega}) e^{j\omega m} d\omega, m \in \mathbf{Z}$$

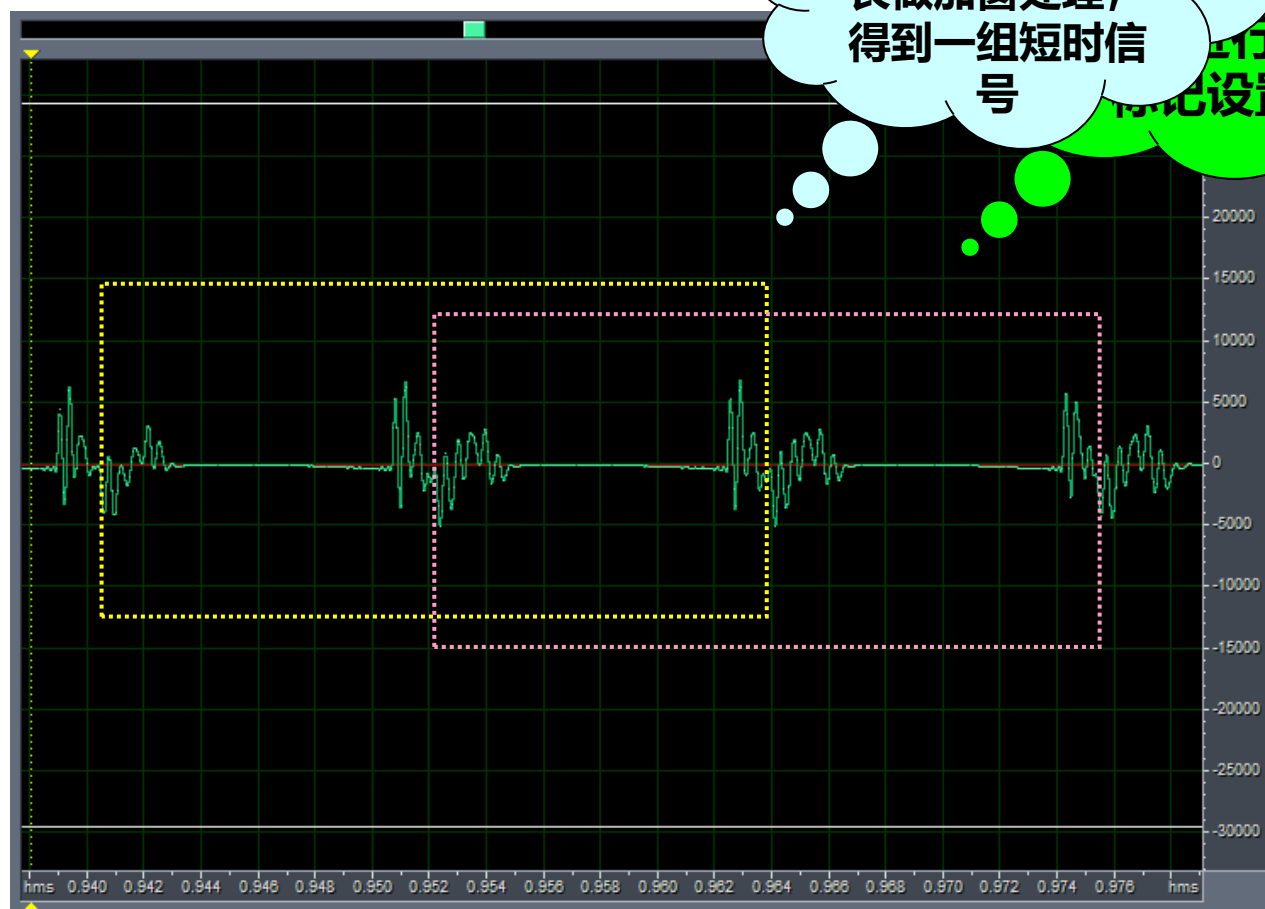
然后叠加 $y_r(m)$ 就能得到原信号

$$y(m) = \sum_{r=-\infty}^{\infty} y_r(m)$$



3:波形平滑

1. 基音同步分析

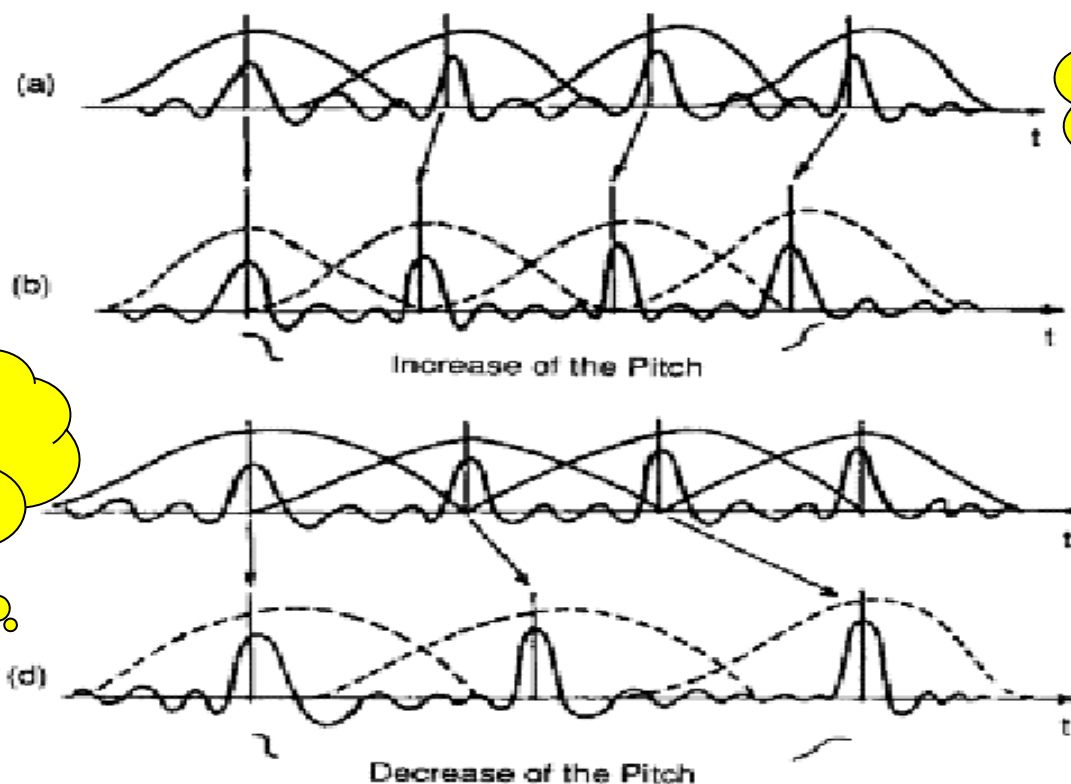


功能
音合
进行同步
标记设置



3:波形平滑

2. 基音同步修改



增加基频

减小基频

分析基音标记
和合成基音标
记未必是一一
对应关系，很
有可能出现一
对多或多对一
的情况



3:波形平滑

3. 基音同步合成

- 谱相等意义下
- 最小均方误差意义下

在一定约束条件下，上述两种方法得到的合成信号
表达形式完全一致，均为

$$y(n) = \sum_q a_q x_{t_m}(n - t_q + t_m)$$



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

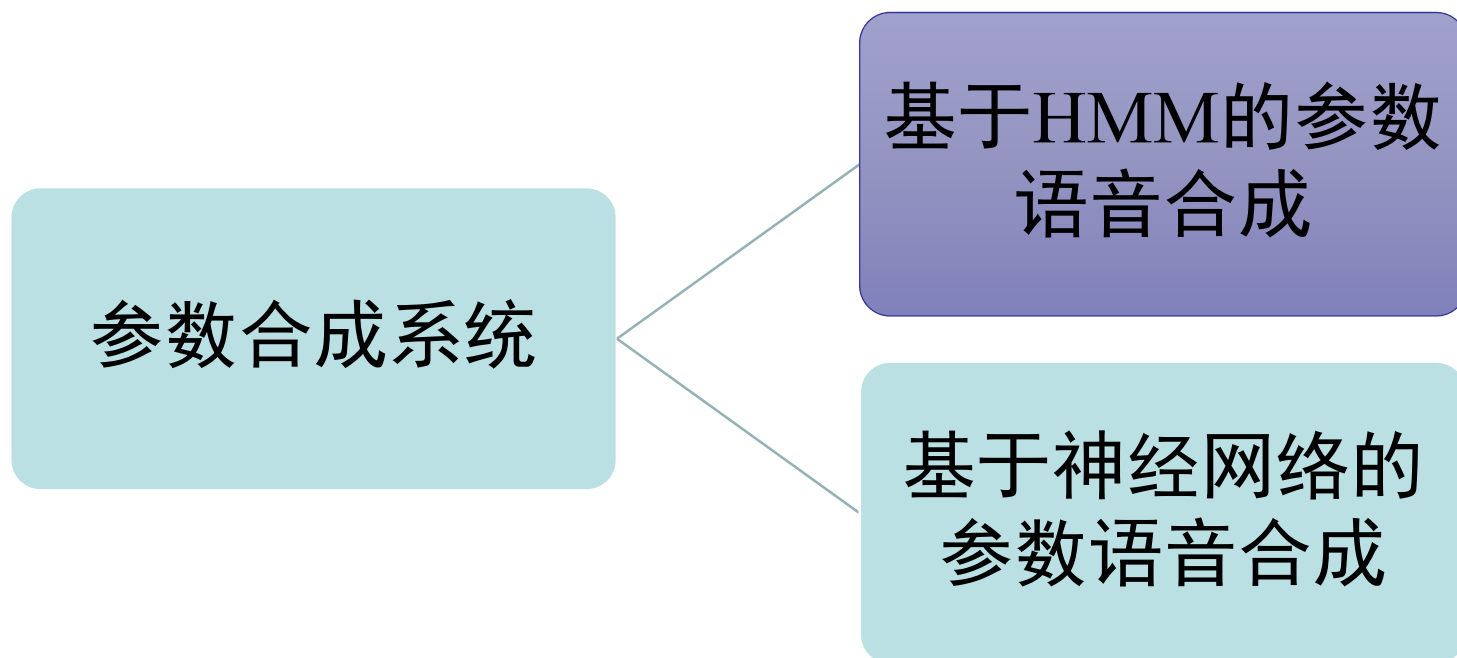
4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



参数合成系统






基于HMM的参数语音合成



参数合成系统

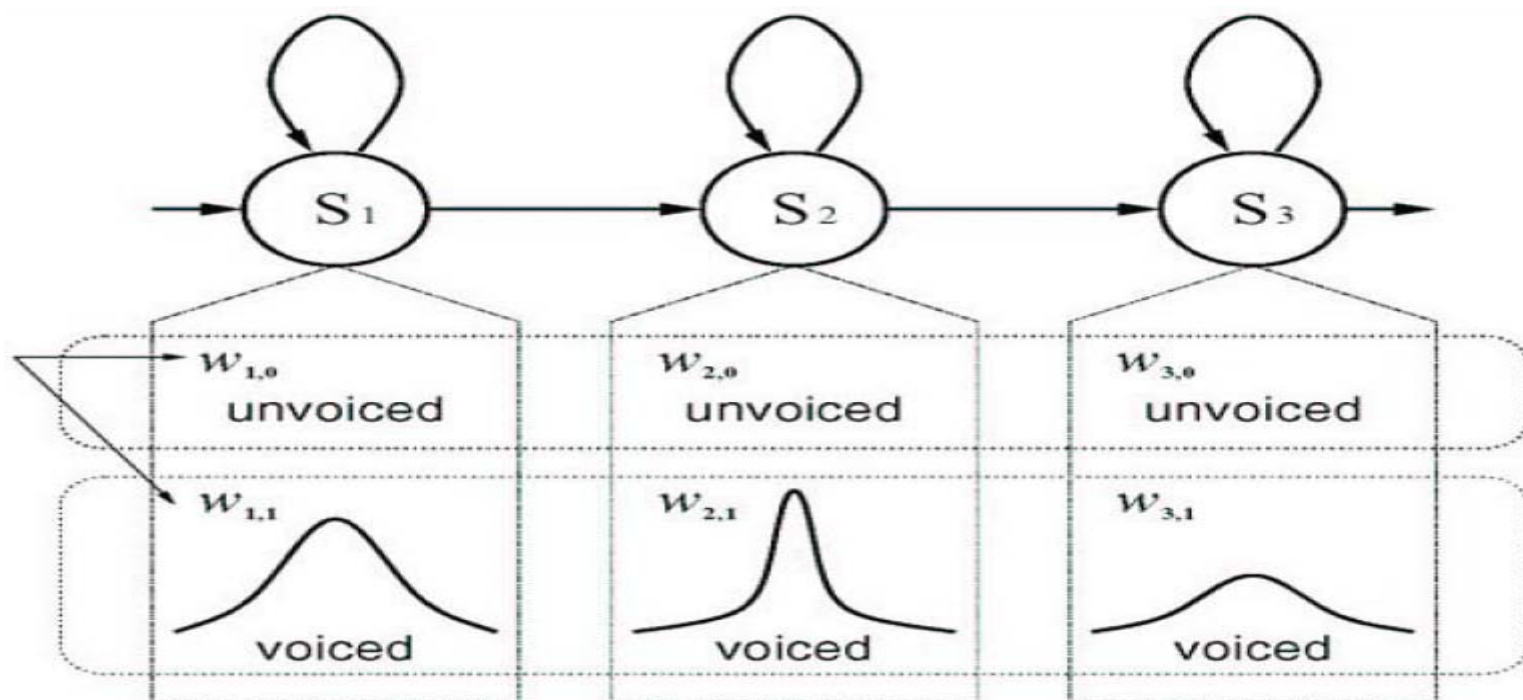
基于HMM的语料库拼接语音合成系统

- 利用HMM目标模型和连接模型来指导单元挑选
- 结合参数训练模型的数学统计模型优势和波形拼接的高音质，相对以前的大语料库技术在自然度上有较大提升
- 自主原发，意义重大
- 优点：拥有明确目标和度量准则，音质好，自然度高，系统搭建自动化程度高
- 缺点：仍然需要很大规模的语料库，计算量较大
- 样例： 



参数合成系统

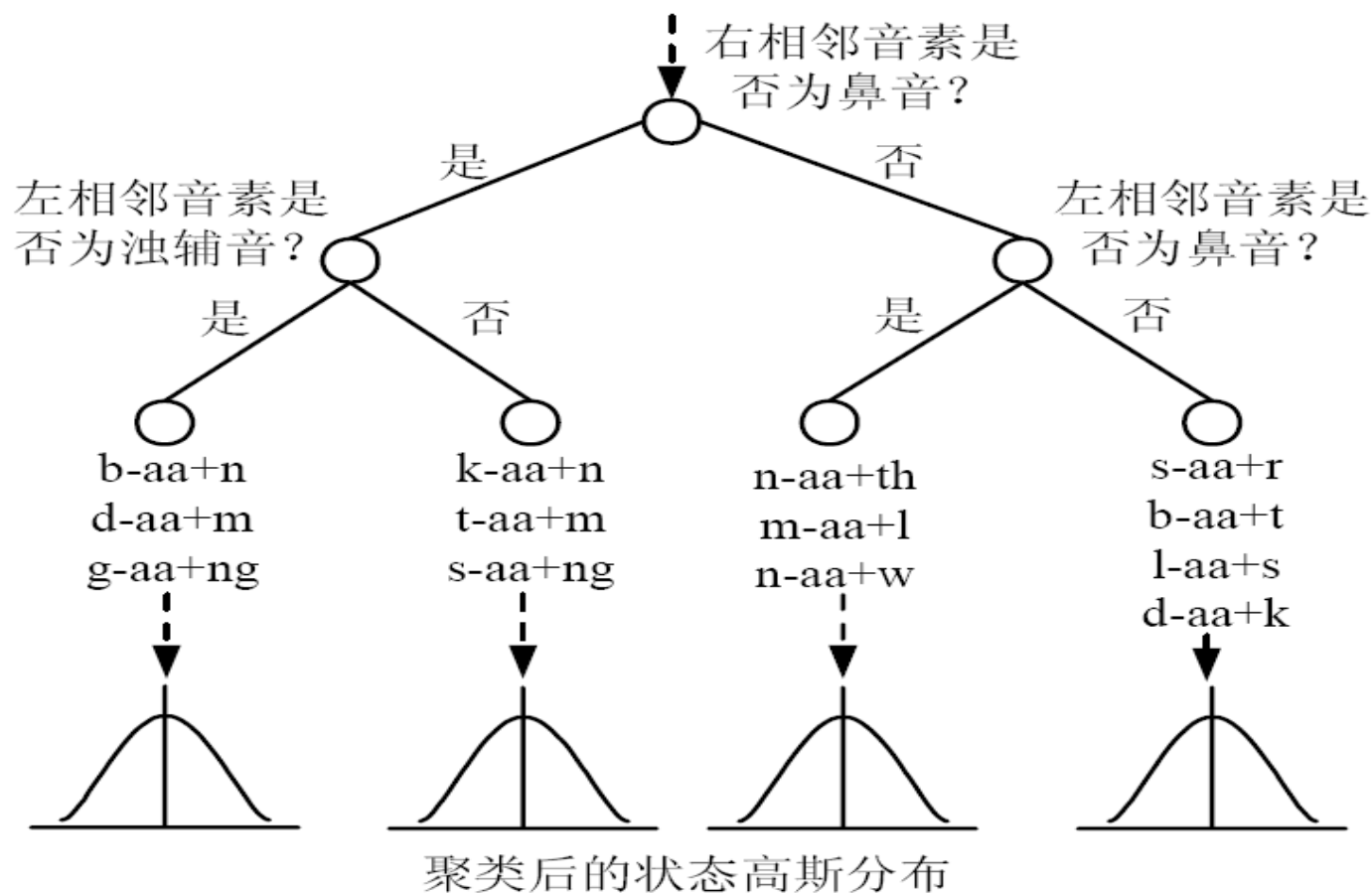
- HMM参数建模
 - 用声学参数针对音素建模
 - 为什么要建模？描述的音素特征变化
 - 隐马尔科夫模型（ Hidden Markov Model -- HMM ）





参数合成系统

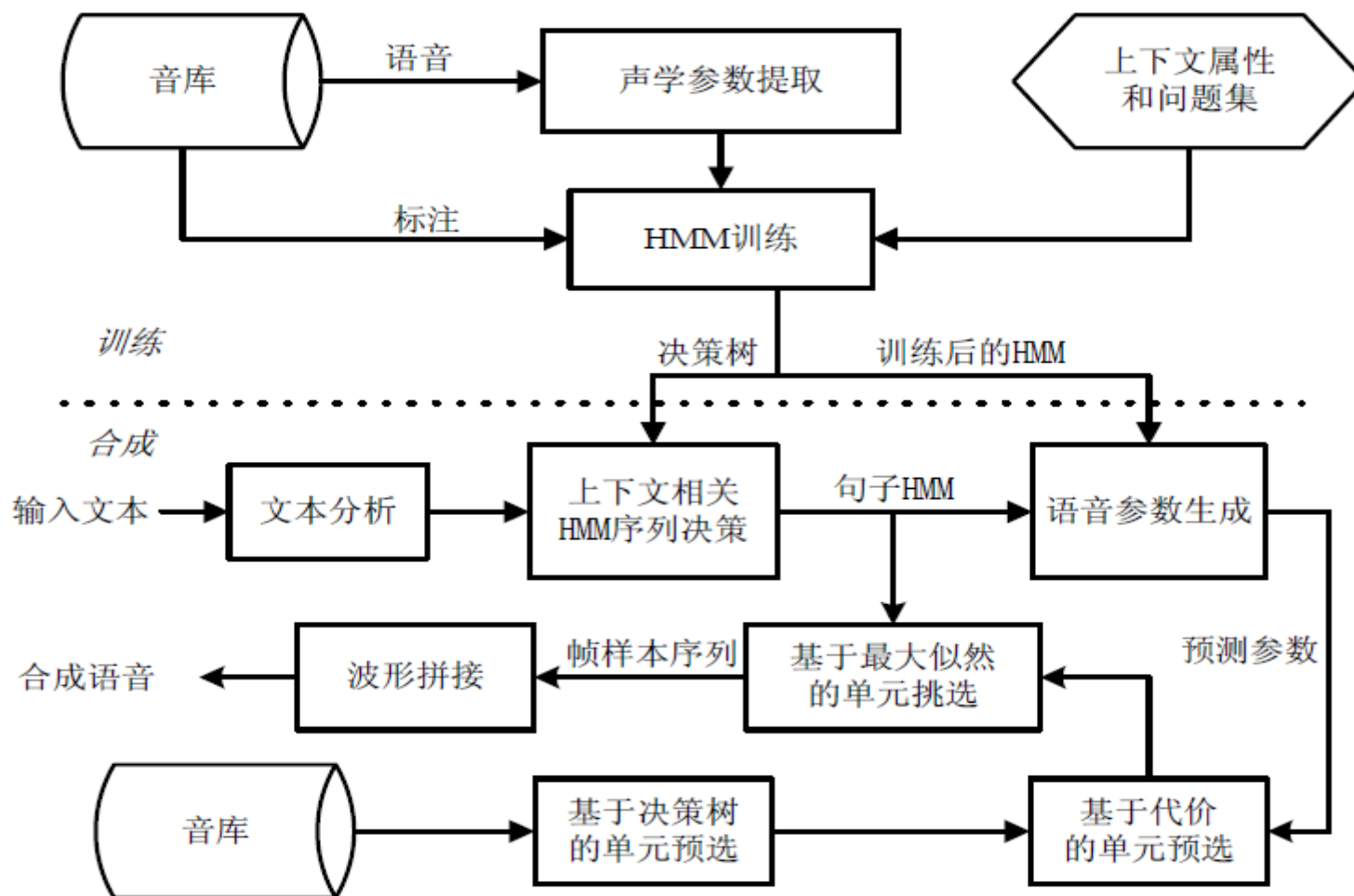
- 决策树模型聚类
 - 有了模型怎么使用？来一句话怎么预知用哪个模型？
 - 基于上下文的信息的决策树聚类





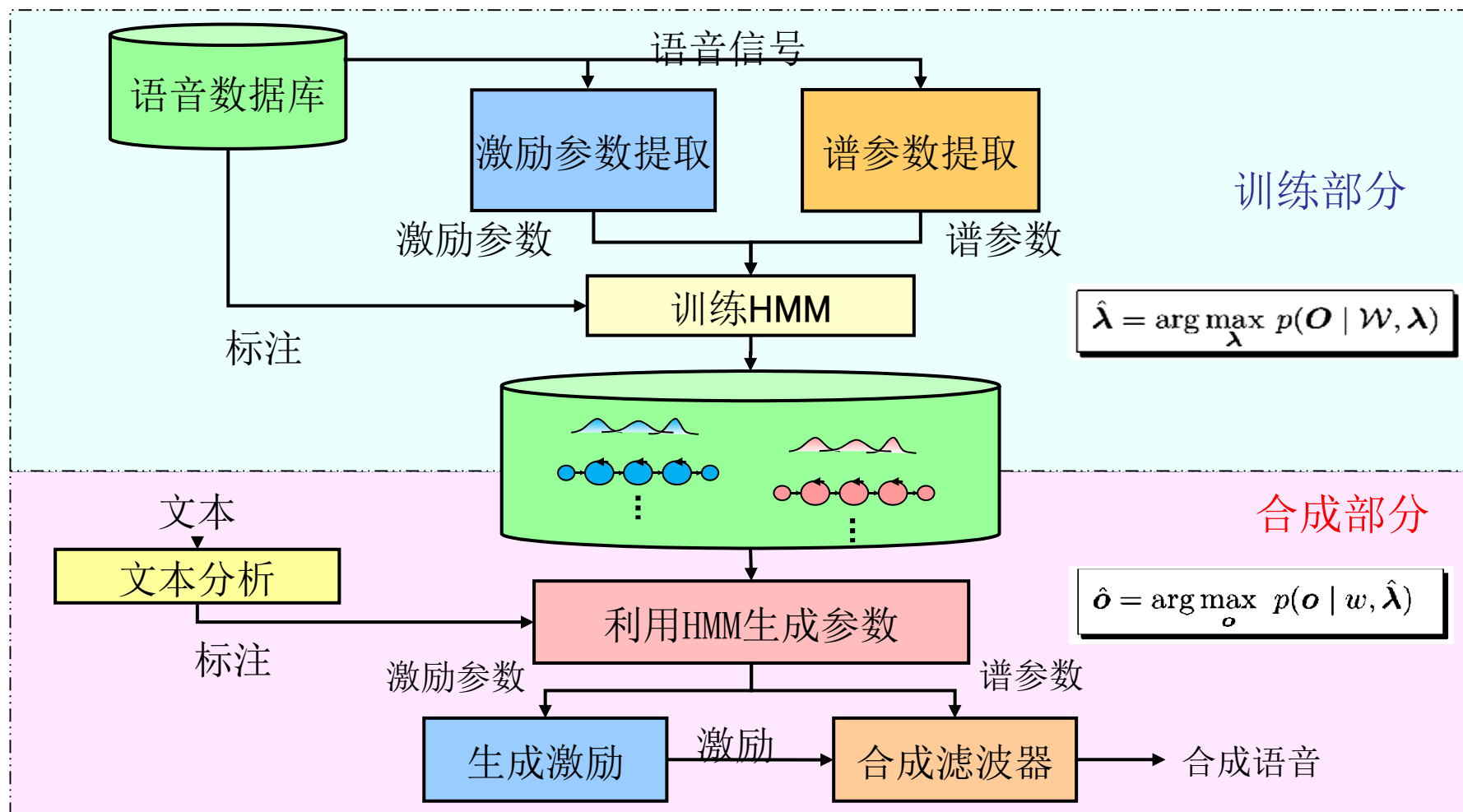
参数合成系统

基于HMM的单元挑选系统结构图





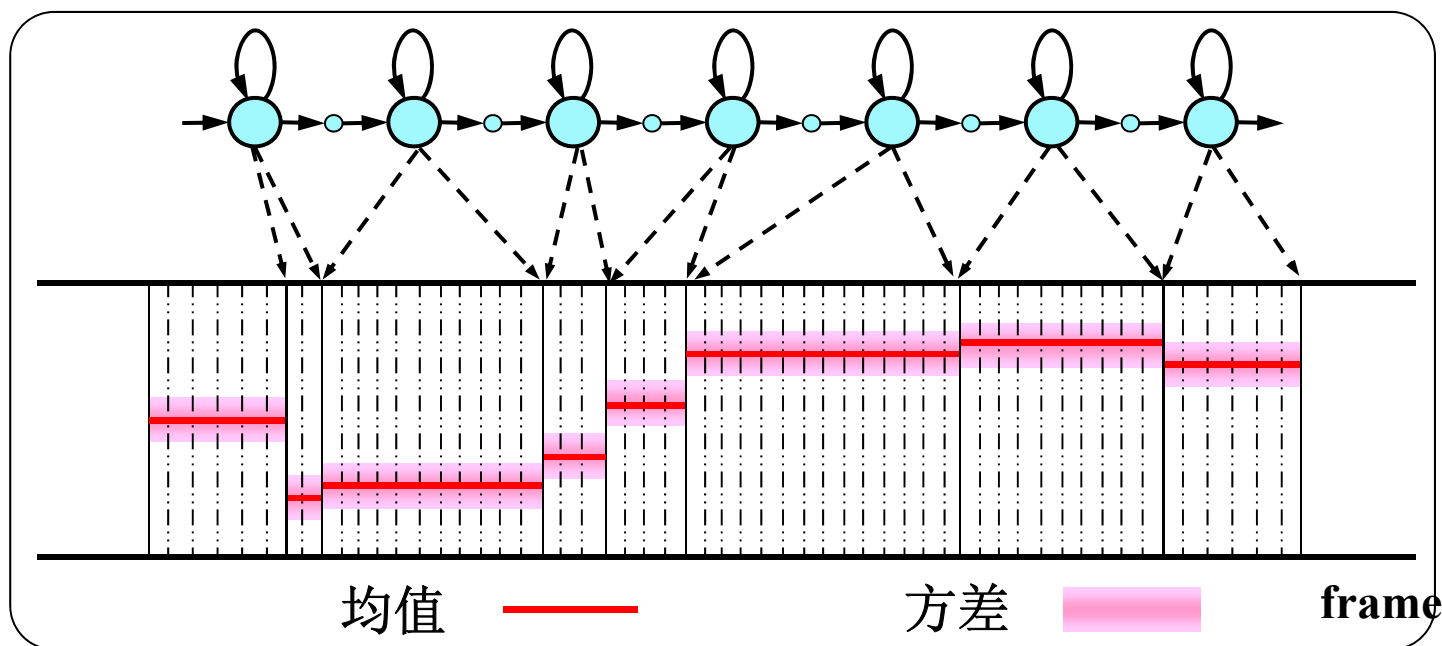
参数合成系统



HMM声学建模示意图



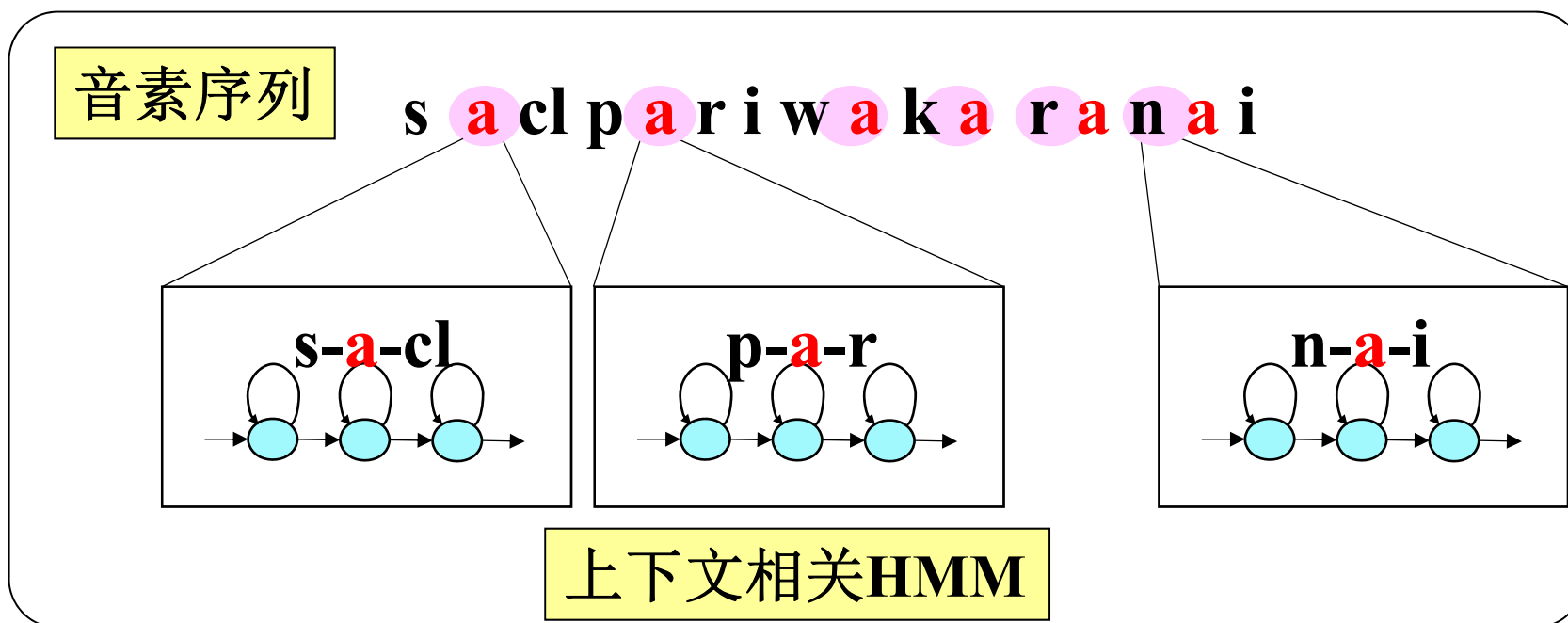
参数合成系统



在已知各个状态时长以及语音参数的均值方差分布后，通过动态参数生成算法确定每帧语音的参数



参数合成系统

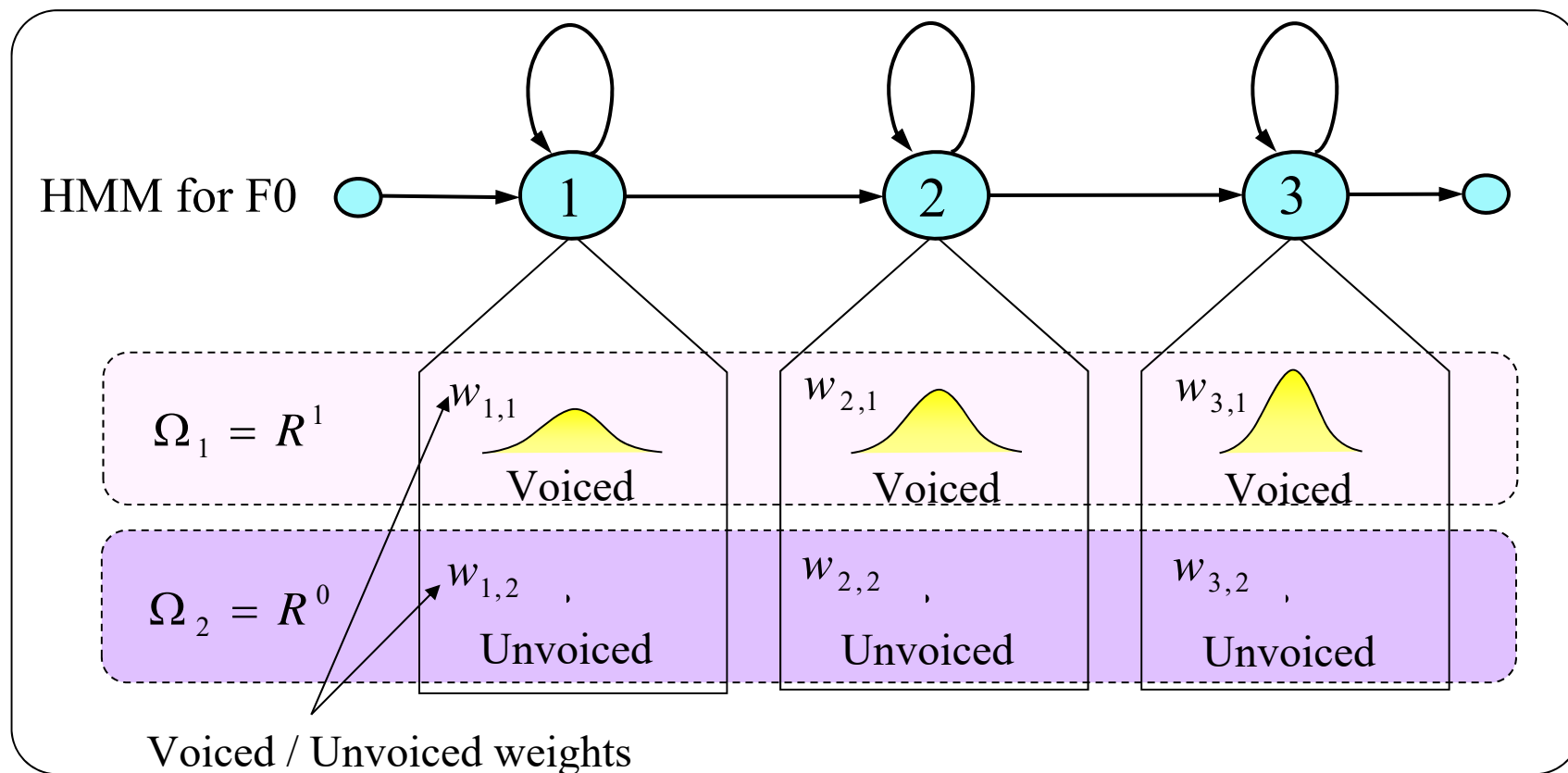


针对协同发音问题，采用上下文相关五音子建模

- 使用聚类后的五音子
- 设计合理的问题集，覆盖丰富的声学信息和语言学信息



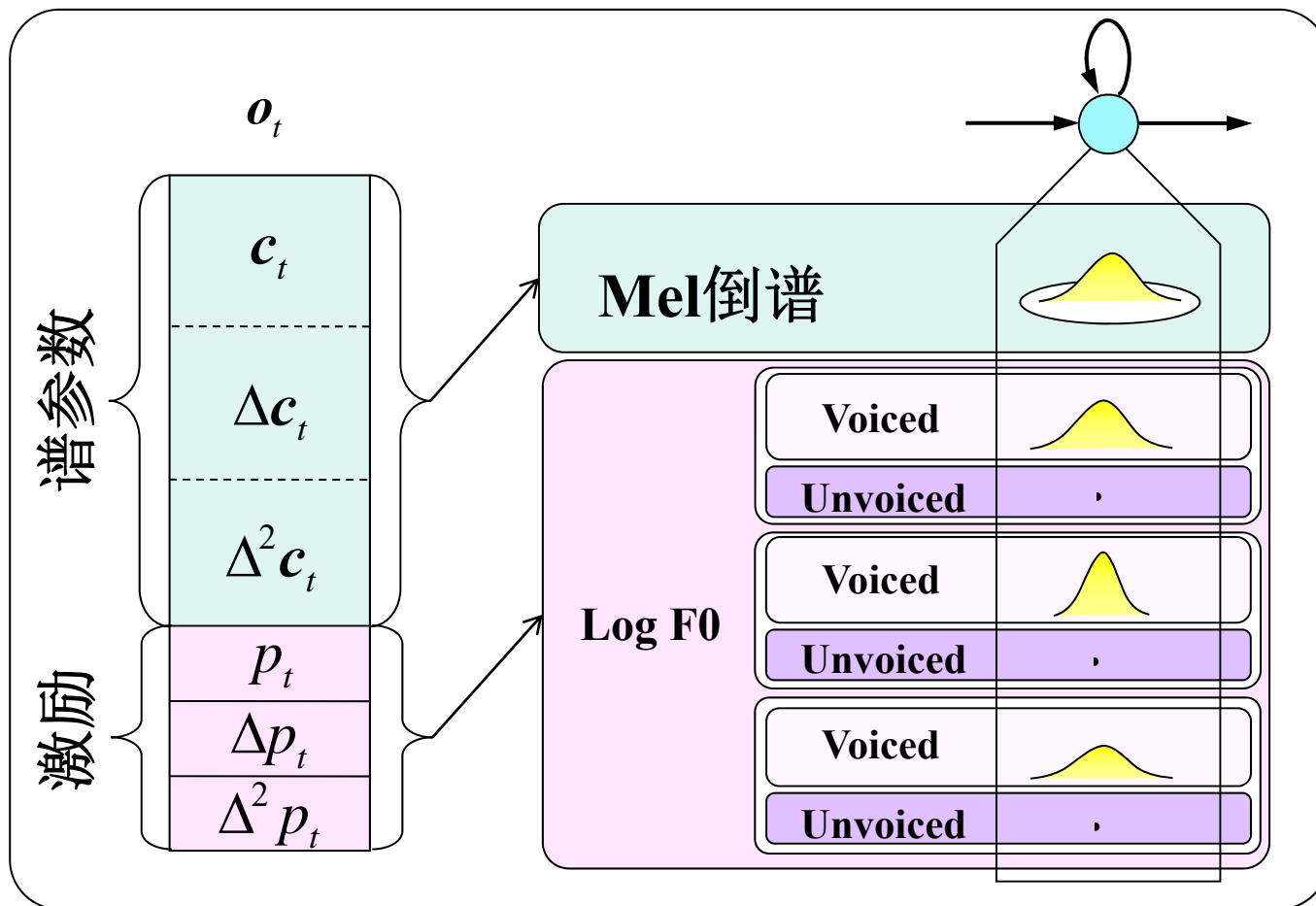
参数合成系统



针对浊音和清音的不同特点，采用MSD描述基频



参数合成系统



HMM状态输出结构示意图



参数合成系统

声码器 (**vocoder**):

声码器的主要作用是基于声学模型生成的声学特征来合成最后声音的波形图

目前语音合成中常见声码器有两种：参数声码器、神经声码器

常用的声码器有：

Griffin-Lim; WORLD/STRAIGHT; WaveNet/Parallel
WaveNet; WaveGlow/FloWaveNet; WaveRNN/LPCNet

Vocoder部分更多涉及到的是语音的知识，如信号处理等相关的知识。



参数合成系统

- 小结：
 - 以音素为单位（中文为声韵母），使用HMM（Hidden Markov Model）对自然语流的频谱特征参数进行建模
 - 采用基于决策树的聚类方法对上下文相关模型进行聚类，以提高模型的鲁棒性，得到预测参数
 - 最后生成参数输入合成器，得到合成语音
 - 优点：所需音库规模小，标注精度要求相对降低，自然度高，系统小，灵活度高
 - 缺点：音质相对较差，带有合成器风格
 - 样例：





参数合成系统

参数合成系统

基于HMM的参数
语音合成

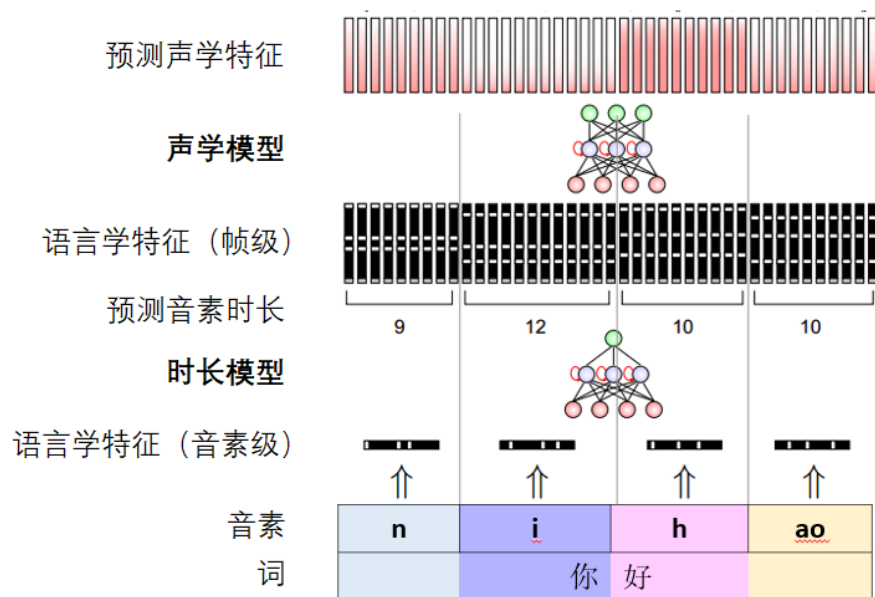
基于神经网络的
参数语音合成



基于神经网络的参数语音合成



基于神经网络的语音合成声学建模



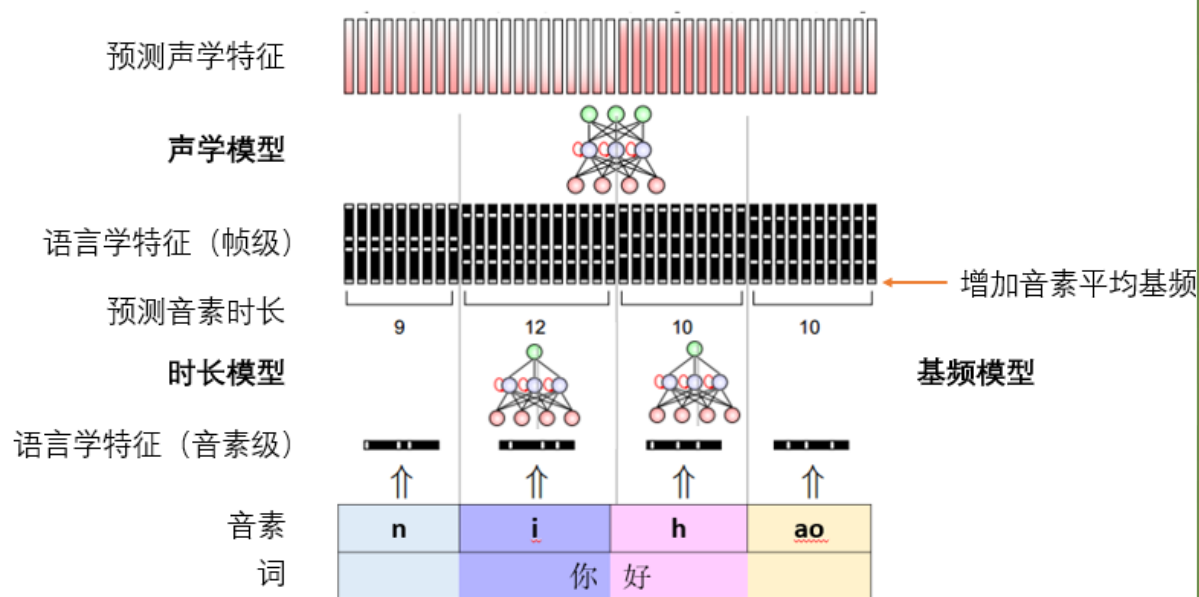
| 声学模型 | MCD(dB) | BAP(dB) | RMSE(Hz) | CORR | VUV |
|---------------------|---------|---------|----------|-------|--------|
| 3*512 DNN | 5.173 | 0.291 | 36.992 | 0.820 | 6.995% |
| 3*512 LSTM | 5.043 | 0.279 | 34.094 | 0.851 | 6.835% |
| 3*256 BLSTM | 4.999 | 0.280 | 33.887 | 0.851 | 6.714% |
| 3*256 (32*5帧) DFSMN | 5.075 | 0.287 | 33.976 | 0.850 | 6.878% |

模型选择

- 研究不同神经网络作为声学模型的性能
- BLSTM作为声学模型性能最优
- 由于LSTM在保证高质量的同时可以使用流式实时合成，因此可作为合成引擎的声学模型



层次化基频建模



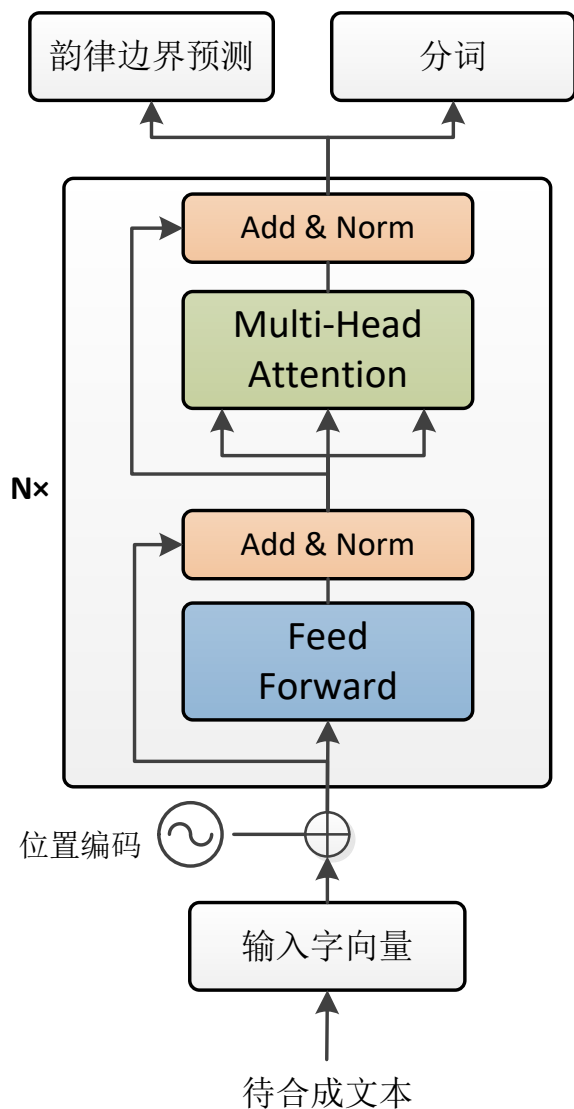
| 声学模型 | RMSE(Hz) | VUV |
|-----------------|----------|--------|
| DNN base | 36.472 | 6.87% |
| DNN + 层级基频 | 33.362 | 6.812% |
| BLSTM base | 33.887 | 6.714% |
| BLSTM + 层级基频 | 33.258 | 6.709% |

层级基频

- 基频预测的精确性
很大程度影响合成
语音的自然度
- 用额外的BLSTM基频
模型预测每个音素
的平均基频
- 声学模型的输入特
征加上音素级基频
，帧级基频的预测
误差有所下降



基于自注意力的语音合成韵律边界预测



□ 端到端

- 避免前端分词及词性标注的影响

□ Self-attention (相对RNN)

- 高度并行化
- 直接对任意两个字间的依赖关系建模

□ 多任务

- 引入词级别信息
- 可用易获取的分词数据预训练

| | 韵律词(F1) | 韵律短语(F1) |
|----------------|---------|----------|
| BLSTM-CRF | 92.99 | 81.69 |
| SELF-ATTENTION | 93.65 | 83.89 |



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



端到端语音合成系统

Tacotron、Tacotron2:

基于RNN结构的seq2seq模型，声学参数预测与声码器分开建模

Transformer TTS:

基于self-attention结构的seq2seq模型，声学参数预测的端到端

ClariNet:

ClariNet 是全卷积模型，训练速度比起基于循环神经网络（RNN）的模型要快 10 倍以上



端到端语音合成

Tacotron2是一个从字符序列生成幅度谱图的seq2seq架构，它仅用输入数据训练出一个单一的神经网络，用于替换语言学和声学特征的生成模块，从而简化了传统语音合成的流水线。

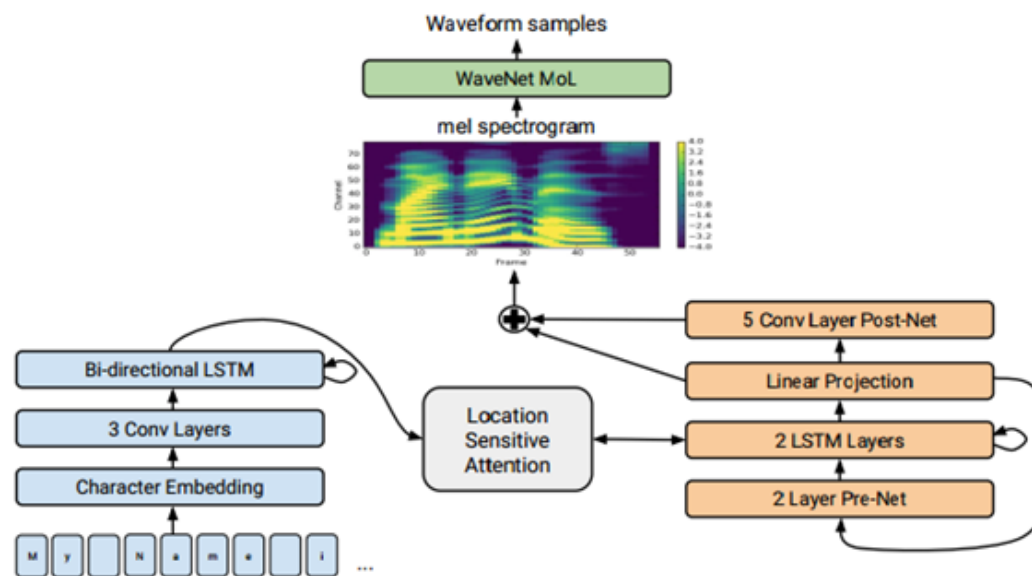
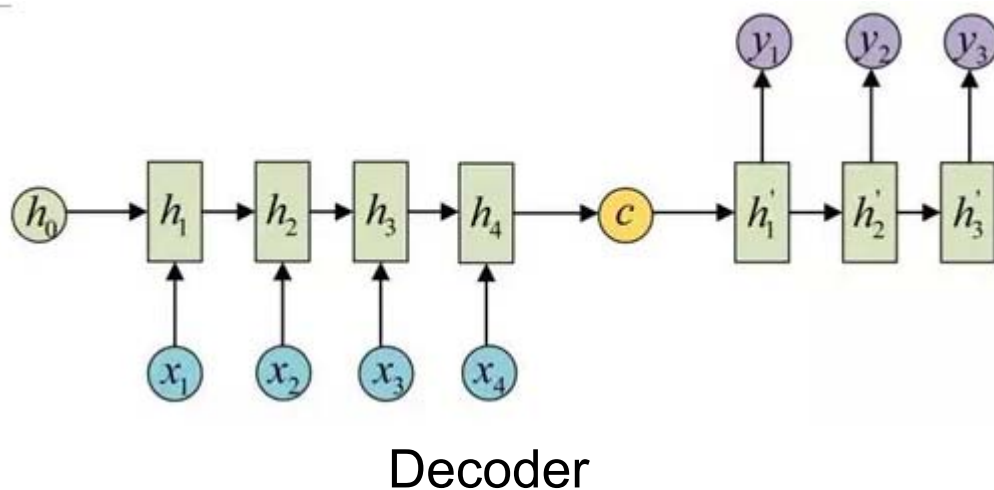
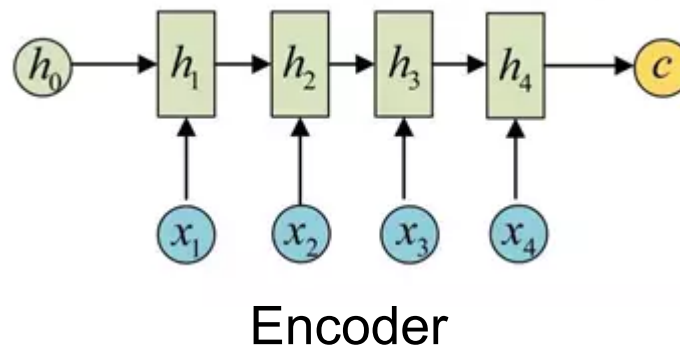


Fig. 1. Block diagram of system architecture.



Seq2seq模型

encoder负责将输入序列压缩成指定长度的向量，这个向量就可以看成是这个序列的语义，这个过程称为编码



decoder则负责根据语义向量生成指定的序列，这个过程也称为解码



端到端语音合成

Tacotron2性能

Tacotron2系统显著优于所有其他TTS系统，其结果可以与标定真实语音相媲美

| Name | MOS |
|-------------------------|-------------------------------------|
| Parametric | 3.492 ± 0.096 |
| Tacotron (Griffin-Lim) | 4.001 ± 0.087 |
| Concatenative | 4.166 ± 0.091 |
| WaveNet (Linguistic) | 4.341 ± 0.051 |
| Ground Truth | 4.582 ± 0.053 |
| Tacotron 2 (this paper) | 4.526 ± 0.066 |

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals for various systems.



大纲

1、语音合成系统概述

2、语音合成系统的前端

3、拼接合成系统

4、参数合成系统

5、端到端语音合成

6、语音合成效果评测



语音合成质量评价



语音合成质量评价

测听与合成

- 效果测听是评判合成系统好坏的硬性指标
- 常用测听项目
 - 音质
 - 自然度
 - 相似度
- 主观打分标准，（mean opinion score, MOS）

| MOS分 | 主观意见 |
|------|---------------|
| 5分 | 优，察觉不到任何不自然 |
| 4分 | 良，刚察觉若干不自然 |
| 3分 | 可，能察觉不自然但可以接受 |
| 2分 | 差，明显察觉但可忍受 |
| 1分 | 坏，不可忍受 |



语音合成质量评价

测听与合成

— 音质测听注意事项

- 对音质由技术路线主导，但敏感度因人而异，主观好恶
 - 16K原始录音音质可打5分
 - 16k原始分析合成可到4分
 - 波形拼接合成音质可超4分
 - 参数合成系统音质在3分附近
- 尽量减少自然度上的错误对音质打分的影响
- 一般测听要求
 - 黑盒：防止惯性打分
 - 0.5分间隔：提高一致性
 - 测听数量不能少，要有覆盖率和代表性
 - 一只好耳机，包住耳朵，提高音量
 - 其实，5分很高，2分很低





测听与合成

– 自然度测听注意事项

- 同样是主观打分，个人标准看待
 - 说话人原始录音也只能接近5分
 - 参数合成较为流畅，相对平淡
 - 拼接合成存在不稳定性，波动较大
- 自然度测听强调对不自然处的**扣分**
- 同样尽量减少不同音质对自然度打分的影响
- 一般测听要求
 - 黑盒：防止惯性打分
 - 0.5分间隔：提高一致性
 - 保证一定数据量，如果数量很多，可以分批测听
 - 5分太高，2分很丢人





测听与合成

– 相似度测听注意事项

- 一般会提供目标人的录音作参照
- 重点考察音色，兼顾基频，时长，口音
- 一般测听要求
 - 黑盒不重要
 - 0.5分间隔：提高一致性





测听与合成

— 偏向性测听注意事项

- 在两个较为接近的效果中取舍
- 测听要求
 - 一定要黑盒！
 - 可以用黑盒工具，固定0， 1打分
 - 偏向性选择只能选一个
 - 在特别说明时，对难以区分的，可以同时选或不选



语音合成demo

男声



真人录音



合成声音

女声



合成声音



真人录音



谢谢

