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

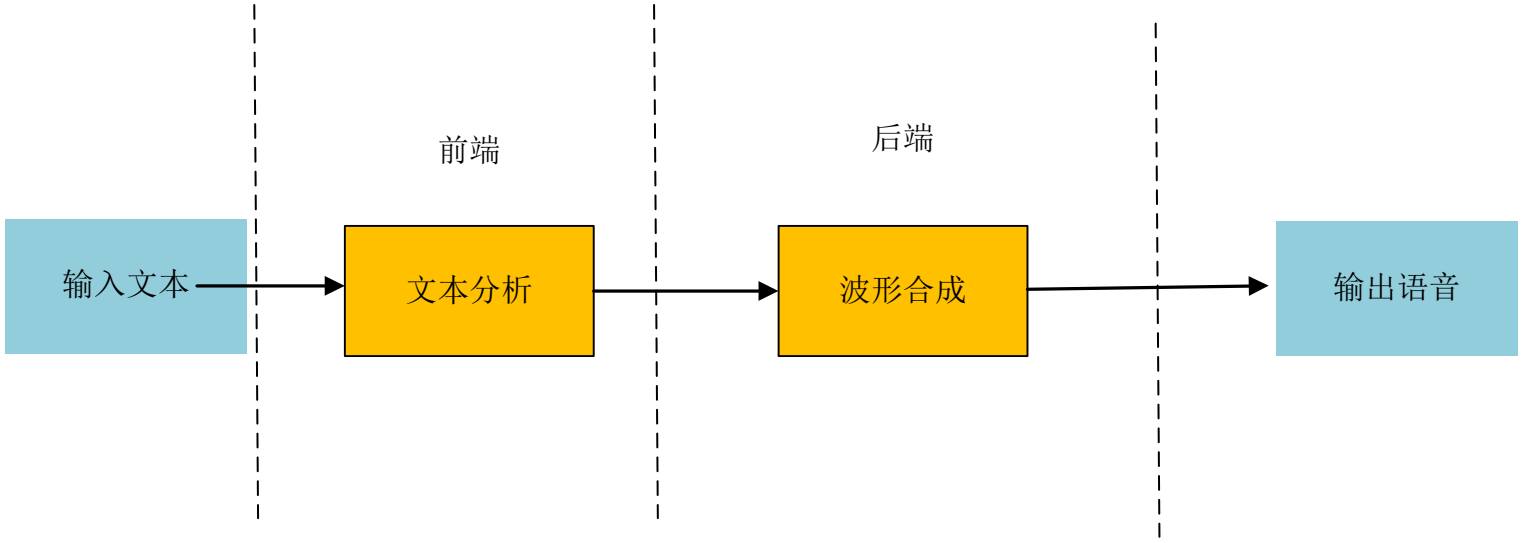
# 语音交互技术
# ——语音合成（一）

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

# 概述

- 语音合成（text to speech，TTS）是一种将文字自动转换为语音信号的技术。语音合成技术涉及声学、语言学、自然语言理解、信号处理、模式识别等多个学科，是信息处理领域的一门前沿技术。



# 语音合成应用


个人助手

车载导航

有声小说

智能语音客服

信息播报

自动应答系统

# 背景与意义

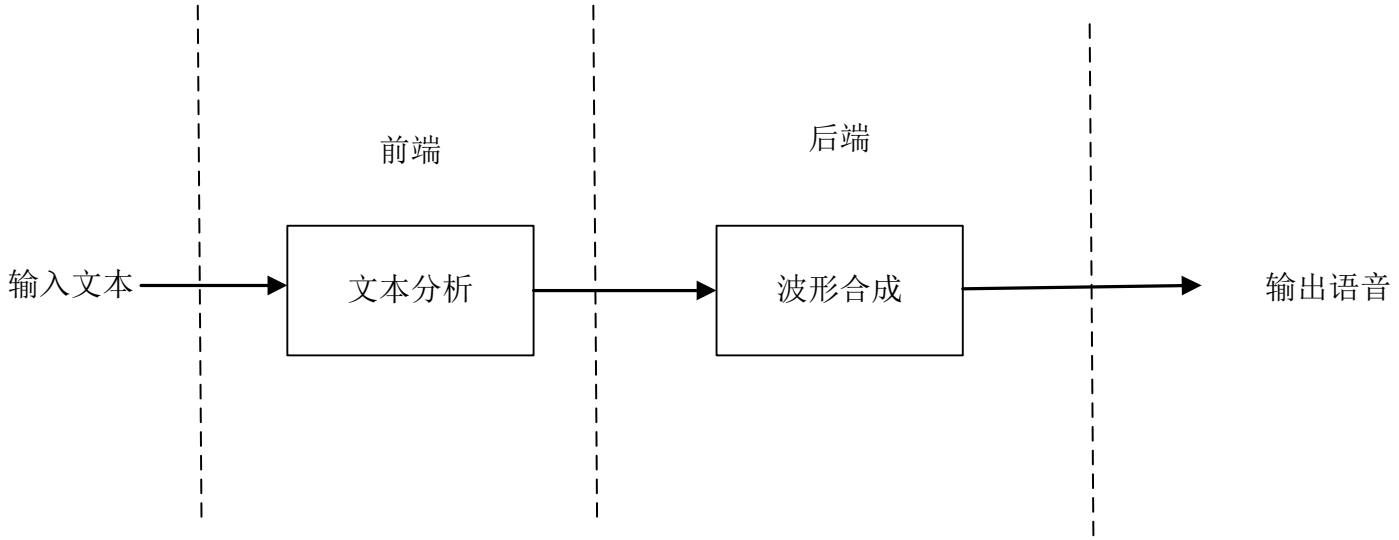
- 目前主流的语音合成方法为数据驱动的方法，包括基于波形拼接的合成方法、基于统计声学建模的合成方法。在统计参数语音合成中，端到端的语音合成方法由于其结构相对简单，需要更少的专家知识，以及独立于语言等特点，成为当前的热点研究内容。

# 语音合成类型

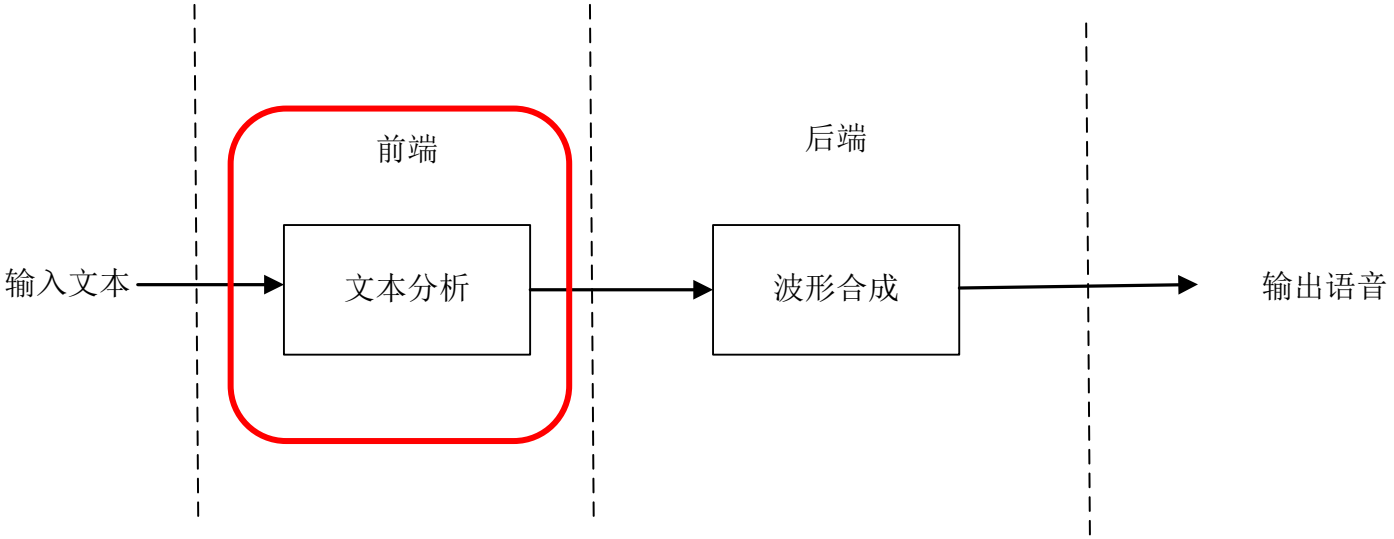
- 管道式语音合成
- 端到端语音合成

# 管道式语音合成

- 语音合成系统主要可以分为文本分析模块、韵律处理模块和声学处理模块。
- 其中文本分析模块可以视为系统的前端，而韵律处理模块和声学处理模块则视为系统的后端。

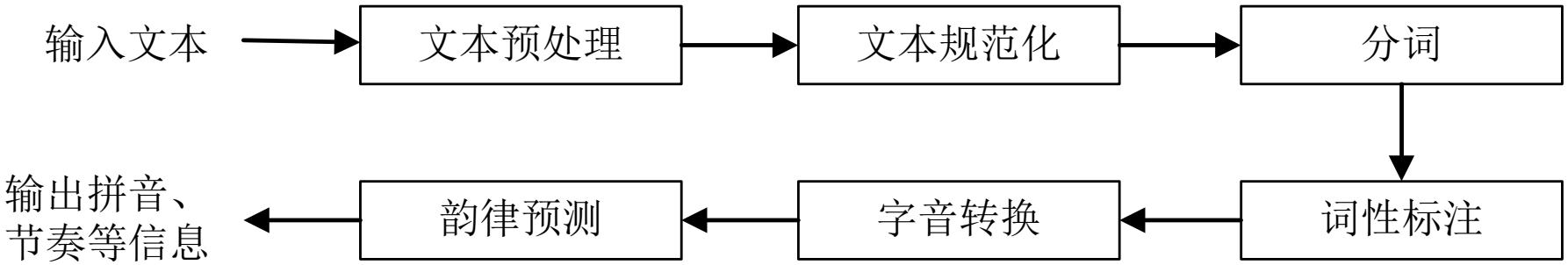


# 管道式语音合成



# 文本分析处理流程

- 文本分析模块是语音合成系统的前端，它的主要任务是对输入的任意文本进行分析，输出尽可能多的语言学信息，如拼音、节奏等，为后端的语音合成器提供必要的信息。
- 对于高自然度的合成系统，文本分析需要给出更详尽的语言学和语音学信息。



# 文本规范化

- 数字的处理
  - 电话号码
  - 日期
  - 时间
  - 钱和货币
  - 符号单位
  - ……
- 特殊符号的处理
  - 冒号“：”、正斜杠“/”、圆点“.”

# 文本规范化方法

- 基于规则的方法
- 基于统计的方法
  - 决策树模型
  - 最大熵方法
  - 信源信道模型
    - 采用统计的方法做文本规范化的工作并不是太多，主要因为要处理的特殊符号与一般的统计语言学有着不同的规律，而且，用于统计训练的特殊符号的标注语料一般也较少，故文本规范化部分多采用基于规则的方法来做。

# 多音字问题

- 汉语中一字多音的现象非常普遍
  - 如：“干”字在“干衣服”中读“gan1”，而在“干重活”中读“gan4”
- 汉语中还有少量多音词，即出现在多字词中的多音字也有多个读音
  - 如：“教授（jiao4shou4 或jiao1shou4）、朝阳（chao2yang2或zhao1yang2）

# 多音字特点

- 大部分的多音字都有多种词性，并且不同的词性对应不同的读音，如上文说的“干”字作形容词时读作“gan1”，作动词时读作“gan4”，因此只要其词性标注正确，就可以根据词性来确定多音字的读音。不过由于多音字存在歧义，其词性较难自动标注准确，如：

1、 枣树/n 长(zhang3)/v 了/u 一点儿/m

2、 裤子/n 长(**zhang3**)/v 了/u 一点儿/m

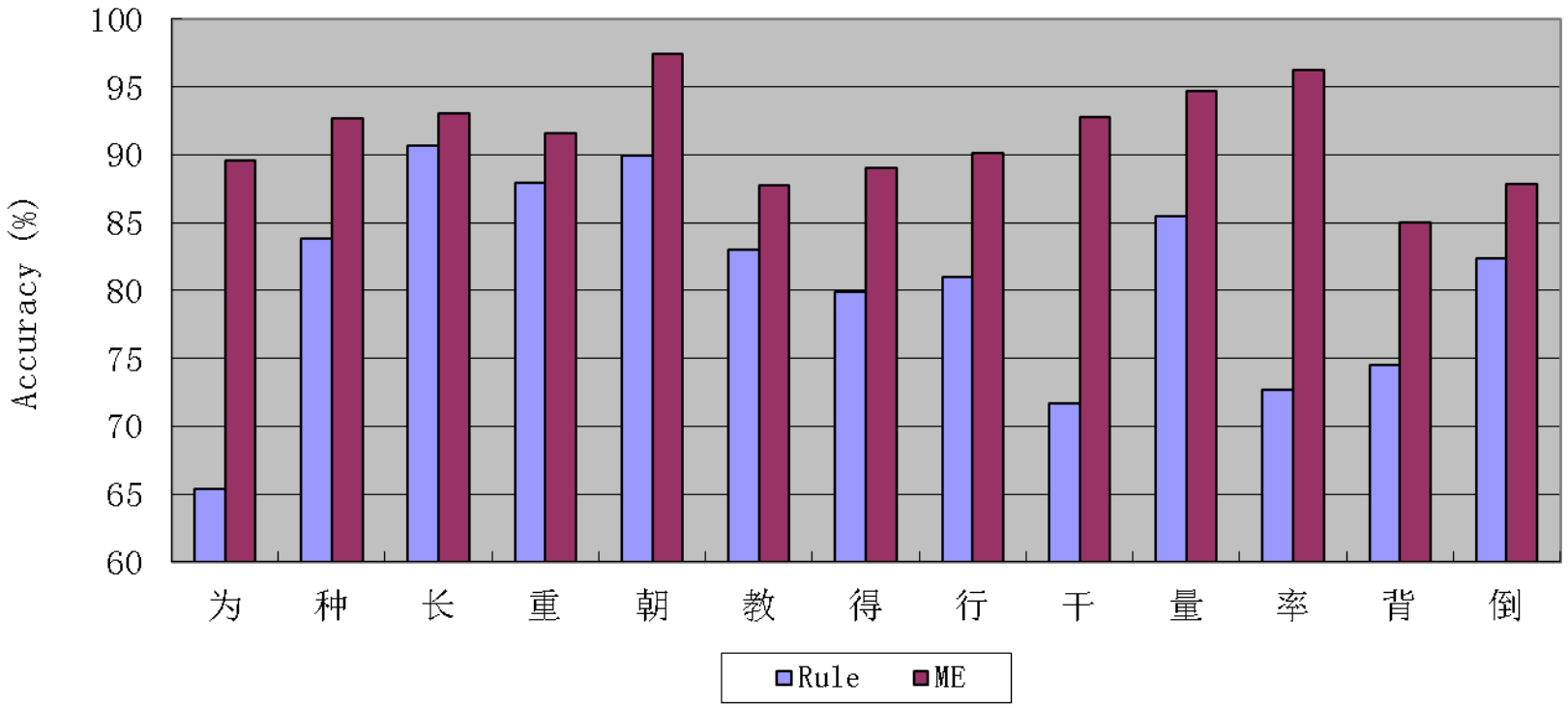
# 多音字分析方法

- 基于手工规则的方法
- 基于机器学习的方法
  - 将多音字消歧看作一个分类问题，其输出是多音字的正确读音，特征空间由多音字附近的其它词组成
  - 特征空间：
    - 1）当前多音字前后4 个词（即左右各2 个词）大小的窗内词（包括词根）的搭配；
    - 2）当前多音字前后4 个词（即左右各2 个词）大小的窗内的词性组合；
    - 3）当前多音字前后4 个词（即左右各2 个词）大小的窗内的词长组合；
  - 常用统计方法：决策树模型、TBL模型、最大熵模型

# 多音字分析方法

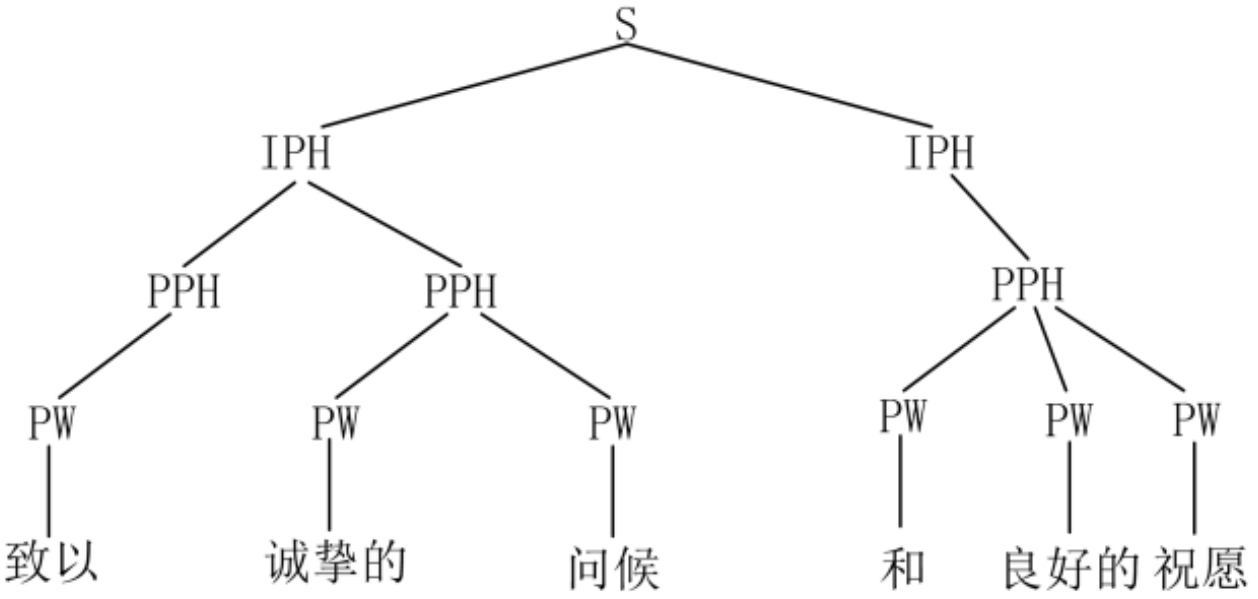
- 特征空间举例：
  - 在山的南面是一条很长的河流。
  - 当前多音字前后4 个词：
    - 一条；很；的；河流
  - 当前多音字前后4 个词性：
    - 量词；副词；助词；名词
  - 当前多音字前后4 个词长：
    - 2；1；1；2
  - ……

# 中文多音字示例



# 韵律层级结构

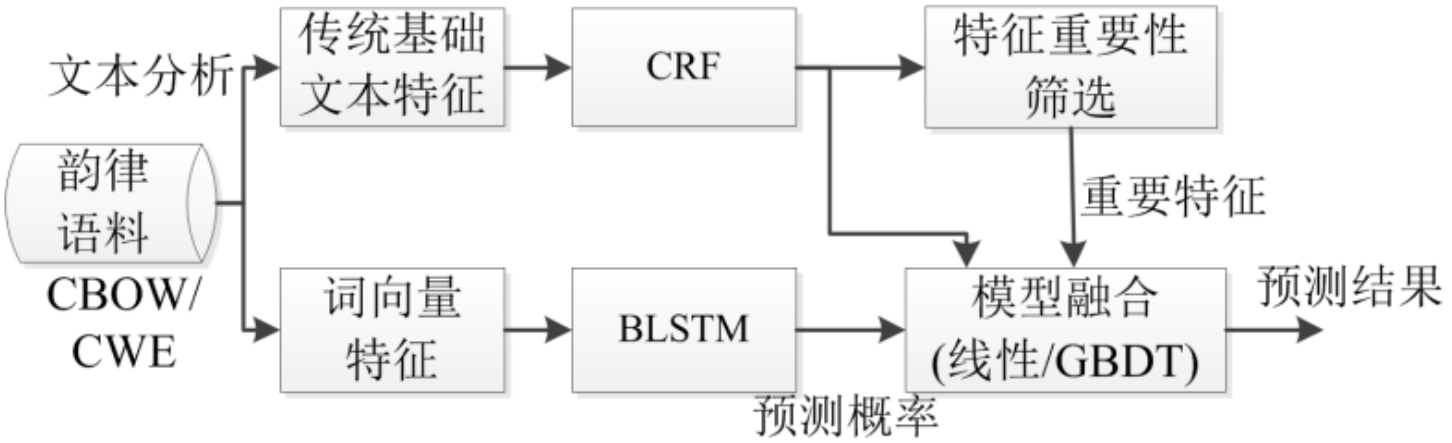
- 句子（ Sentence, S）
- 韵律词（Prosodic Word, PW）
- 韵律短语（Prosodic Phrase, PPH）
- 语调短语（Intonational Phrase, IPH）



# 韵律层级结构

- 韵律词（PW）
  - 韵律词是最小的能够自由运用的语言单位。通俗地讲，韵律词是一组在实际语流中经常在一起发音的音节。
- 韵律短语（PPH）
  - 韵律短语是语法结构的停延和音步的音域展敛可以使用规则控制的可能多音步。我们认为汉语的韵律词决定性的韵律标记是单音步，而韵律短语是在单音步组合的基础上再加上更高层的停延和音步的音域展敛变化。
- 语调短语（IPH）
  - 语调短语是指具有完整的语调，听感上可独立成句的一段发音。也就是将几个韵律短语按照一定的句调模式连接起来，一般对应句法上的句子。

# 韵律节奏预测

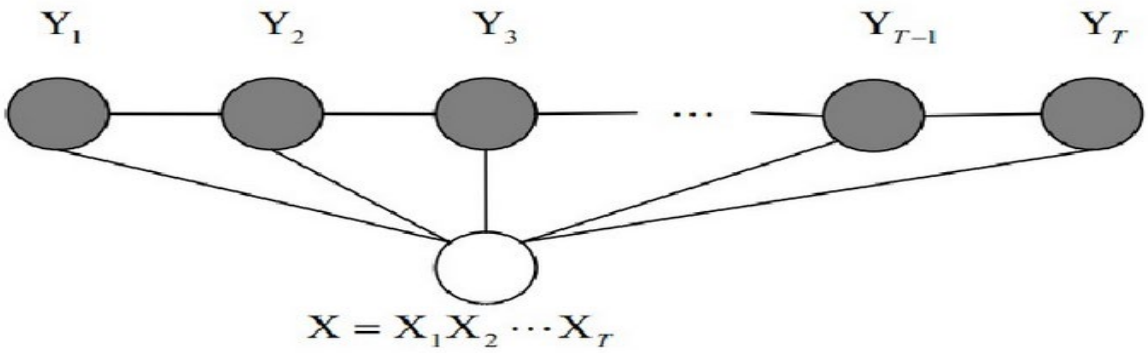



- 传统方法存在的问题
  - ✓ 只考虑人工设计的、启发式的文本特征，如词性、词长等
  - ✓ 单模型的预测结果往往不佳
- 深度学习方法
  - ✓ 使用自动学习的文本特征，如字向量，词向量等
  - ✓ 采用多模型融合进行预测

# 条件随机场：基本概念

## 什么是条件随机场？

- 提出：
  - 条件随机场模型是Lafferty于2001年，在最大熵模型和隐马尔可夫模型的基础上，提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。
- 概念：
  - 条件随机场（Conditional Random Field，CRF）是一种判别式无向图模型。条件随机场试图对多个变量在给定观测值后的条件概率进行建模。具体来说，已知观测序列，和与之对应的标记序列，则条件随机场的目标是构建条件概率模型。

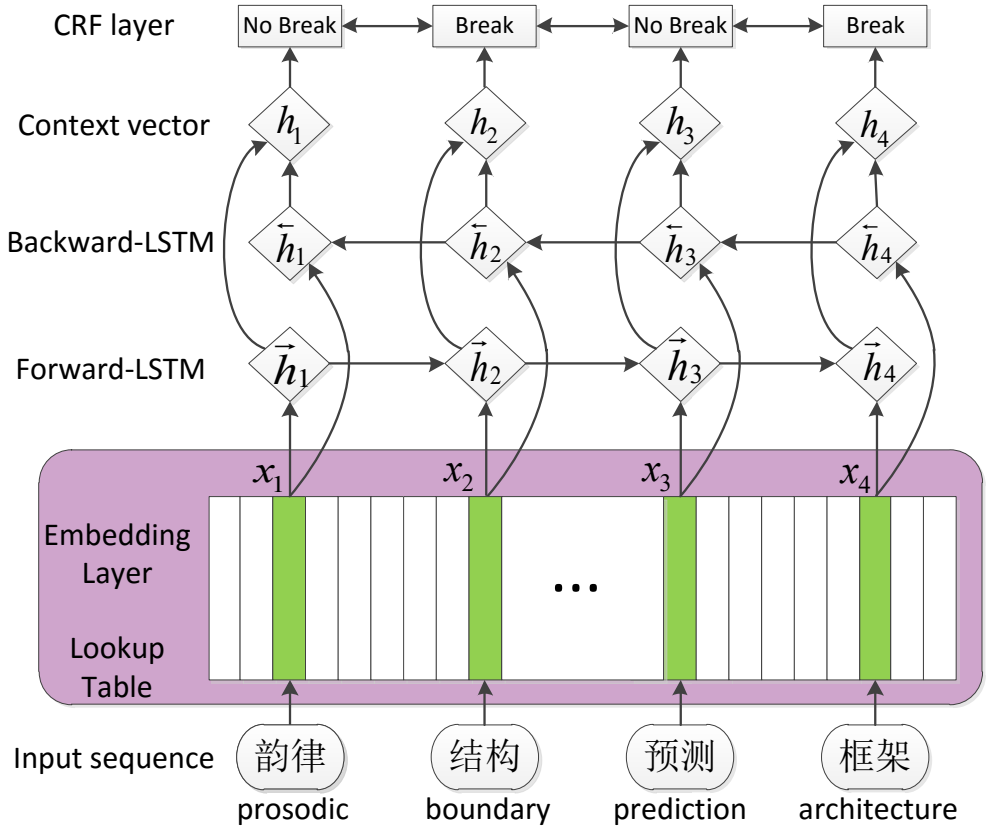


# 条件随机场模型：韵律预测模型中的应用

- 韵律预测模型：
  - 对于韵律节奏预测任务来说，观测数据为语句，标记为相应的韵律节奏停顿序列，具有线性序列结构，标记序列中的变量之间具有相关性，所以变量是结构型变量。
  - 条件随机场预测韵律节奏主要在于采取上下文信息（特征）来进行节奏的预测。
  - 首先人工根据训练语料的规律来制定特征模板，其次根据特征模板，从训练的语料中选取符合模板的特征值来进行预测。
  - 而特征模板主要通过词本身、词性、词长、当前位置到句子两端的距离来制定。

# 基于深度神经网络的节奏预测（ BLSTM-CRF ）

- 相比于深度置信网络，循环神经网络更适合应用于序列标注问题。它被证明在很多自然语言理解的序列标准问题中，都取得了较好的效果。




BLSTM-CRF用于韵律边界预测

文本Embedding层

# 常见方法的节奏预测对比分析

中文数据库上F1值

| 系统 | CRF | BLSTM_CBOW | BLSTM_CEW | LF | GBDT1 | GBDT2 |
|---|---|---|---|---|---|---|
| 韵律词 | 95.52 | 95.79 | 95.90 | 96.19 | 96.43 | **96.79** |
| 韵律短语 | 82.25 | 82.95 | 83.36 | 84.17 | 84.82 | **85.25** |
| 语调短语 | 79.51 | 81.08 | 81.74 | 82.88 | 83.67 | **84.73** |

| 模型 | 输入特征描述 |
|---|---|
| CRF | 词性和长度等传统基础文本特征，详见表 2.2 |
| BLSTM_CBOW | CBOW 模型预训练的字或词向量 |
| BLSTM_CWE | CEW 模型预训练的字或词向量 |
| LF | CRF 和 BLSTM 的概率输出，不包含筛选出的传统重要性特征 |
| GBDT1 | CRF 和 BLSTM 的概率输出，不包含筛选出的传统重要性特征 |
| GBDT2 | CRF 和 BLSTM 的概率输出，并包含筛选出的传统重要性特征 |

# 常见方法的节奏预测对比分析

## 中文数据库上F1值

| Systems | CRF | BL | BC | WB | CC | CA |
|---|---|---|---|---|---|---|
| 韵律词 | 95.46 | 95.60 | 96.01 | 96.26 | 96.49 | **96.67** |
| 韵律短语 | 79.39 | 80.15 | 81.49 | 82.04 | 82.48 | **82.75** |
| 语调短语 | 77.54 | 78.88 | 80.57 | 81.20 | 81.69 | **81.89** |

## 英文数据库上F1值

| Systems | CRF | BL | BC | WB | CC | CA |
|---|---|---|---|---|---|---|
| 语调短语 | 75.24 | 75.78 | 77.44 | 77.79 | 79.13 | **79.52** |

- CRF：传统的文本特征（包含词性、单词长度等），配合CRF用于韵律边界预测
- BL：预训练的词向量，配合BLSTM用于韵律边界预测
- BC:预训练的词向量，配合BLSTM-CRF模型用于韵律边界预测
- WB:增强词嵌入层，使得词向量可以随着模型联合优化，配合BLSTM-CRF模型用于韵律边界预测
- CC:基于注意力机制融合的字、词向量，配合BLSTM-CRF模型
- CA:使用上下文敏感词向量，配合BLSTM-CRF模型用于韵律边界预测

# 常见方法的节奏预测对比分析

中文数据库上F1值

| Systems | CRF | BL | BC | WB | CC | CA |
|---|---|---|---|---|---|---|
| 韵律词 | 95.46 | 95.60 | 96.01 | 96.26 | 96.49 | **96.67** |
| 韵律短语 | 79.39 | 80.15 | 81.49 | 82.04 | 82.48 | **82.75** |
| 语调短语 | 77.54 | 78.88 | 80.57 | 81.20 | 81.69 | **81.89** |

英文数据库上F1值

| Systems | CRF | BL | BC | WB | CC | CA |
|---|---|---|---|---|---|---|
| 语调短语 | 75.24 | 75.78 | 77.44 | 77.79 | 79.13 | **79.52** |

- 和CRF系统相比，使用BSTM-CRF，能综合利用BLSTM和CRF方法的优势，其中BLSTM层能够很好的刻画序列的上下文依赖关系，CRF能够实现句子级别的联合解码。
- 引入上下文敏感的词向量，能够进一步提升实验结果。
- 所提方法与语言无关，在英文预料上取得了和中文预料上相似的结果

# 时长预测

## 时长模型

- 时长是指某个发音的持续时间，时长模型为输入待合成文本中每个音素分配所需要持续的时间长度。

## 具体步骤（以英文为例）

- 步骤1：字素转音素（Grapheme to Phoneme, G2P）
  - Input - "It was earky spring"
  - Output - [IH1, T, ., W, AA1, Z, ., ER1, L, IY0, ., S, P, R, IH1, NG,. ]
- 步骤2：音素时长
  - Input - [IH1, T, ., W, AA1, Z, ., ER1, L, IY0, ., S, P, R, IH1, NG,.]
  - Output - [IH1 (0.1s), T(0.05s),. (0.01s),...]

# 时长预测

## 时长模型

- 时长是指某个发音的持续时间，时长模型为输入待合成文本中每个音素分配所需要持续的时间长度。

## 具体步骤（以中文为例）

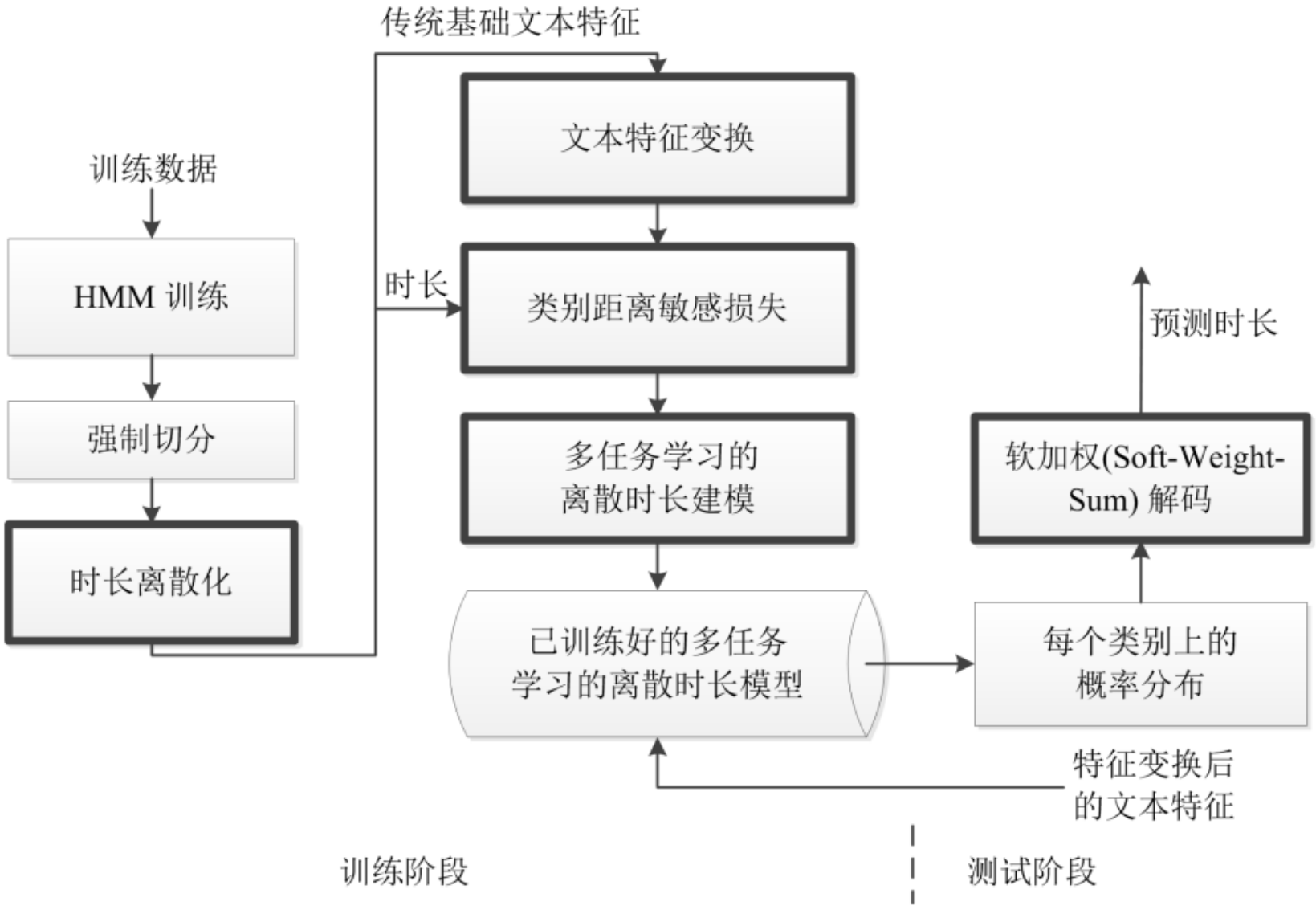
- 步骤1：汉字转音素

  •输入-【她一言不发】
  •输出-【t a1 i4 ian2 b u4 f a1】

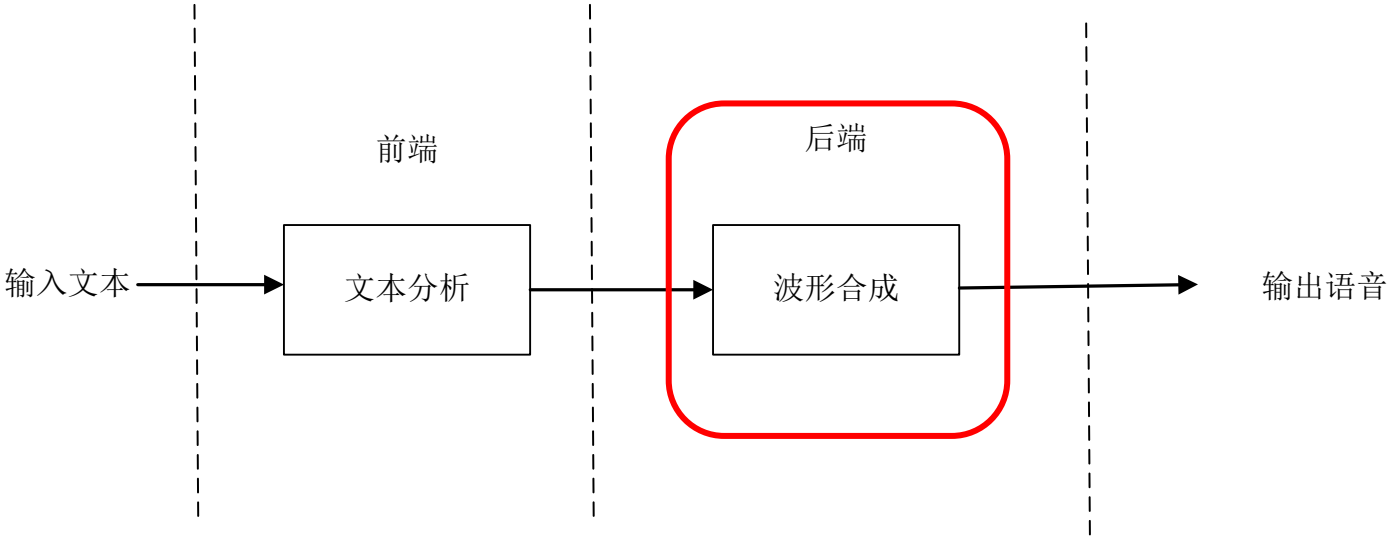
- 步骤2：音素时长

  • 输入-【t a1 i4 ian2 b u4 f a1】
  • 输出-【t(0.02s) a1(0.2s) i4 (0.1s)…】

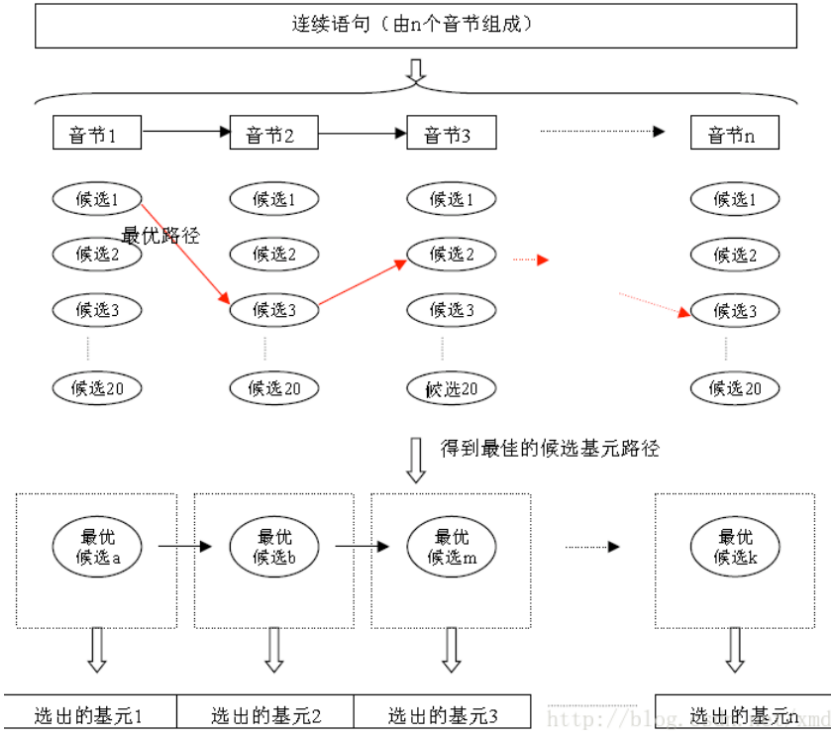
# 时长预测训练流程

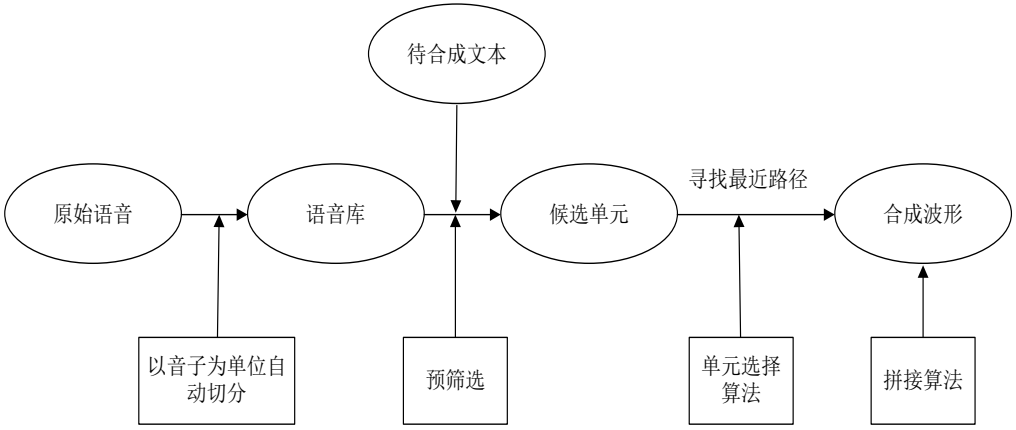

传统基础文本特征
文本特征变换
训练数据
HMM 训练
时长
类别距离敏感损失
预测时长
强制切分
多任务学习的
离散时长建模
软加权(Soft-Weight-
Sum) 解码
时长离散化
已训练好的多任务
学习的离散时长模型
每个类别上的
概率分布
特征变换后
的文本特征
训练阶段
测试阶段

# 管道式语音合成



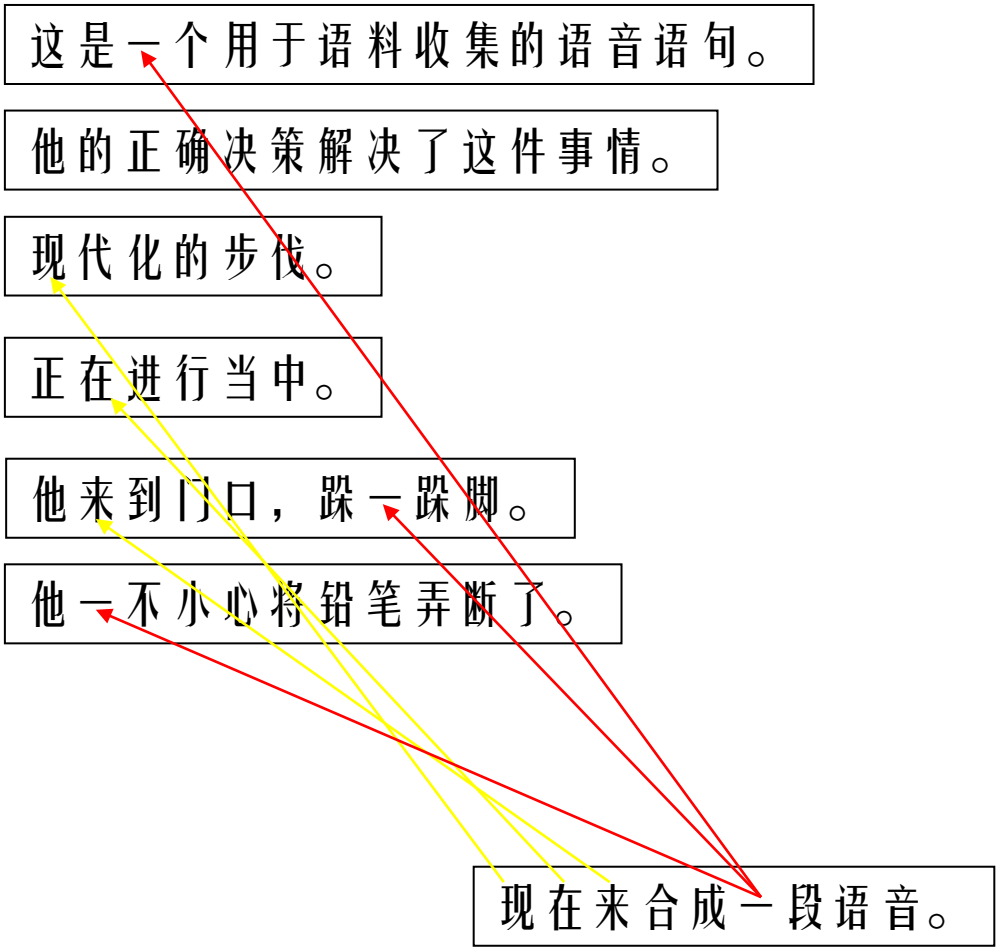
# 语音合成的波形拼接方法




基于波形拼接的语音合成框架

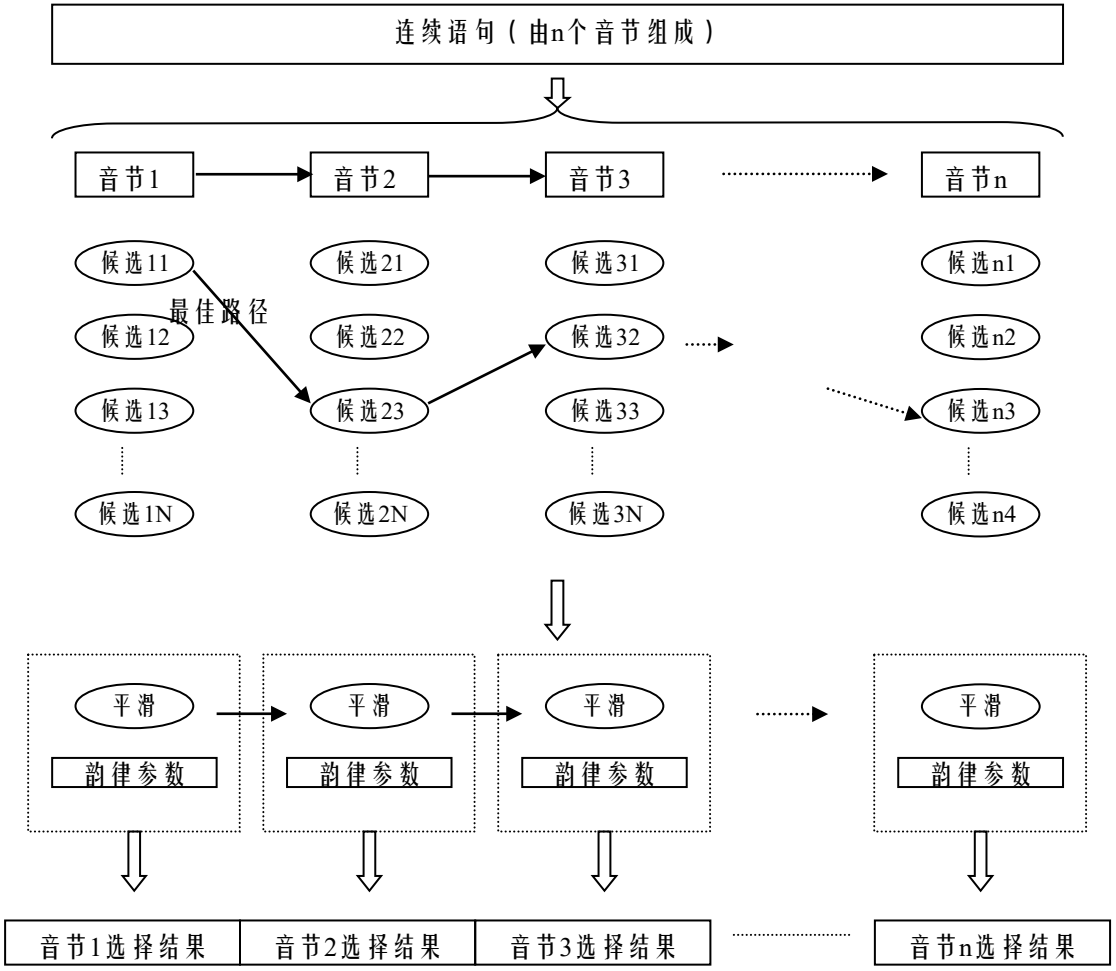
# 语音合成的波形拼接方法



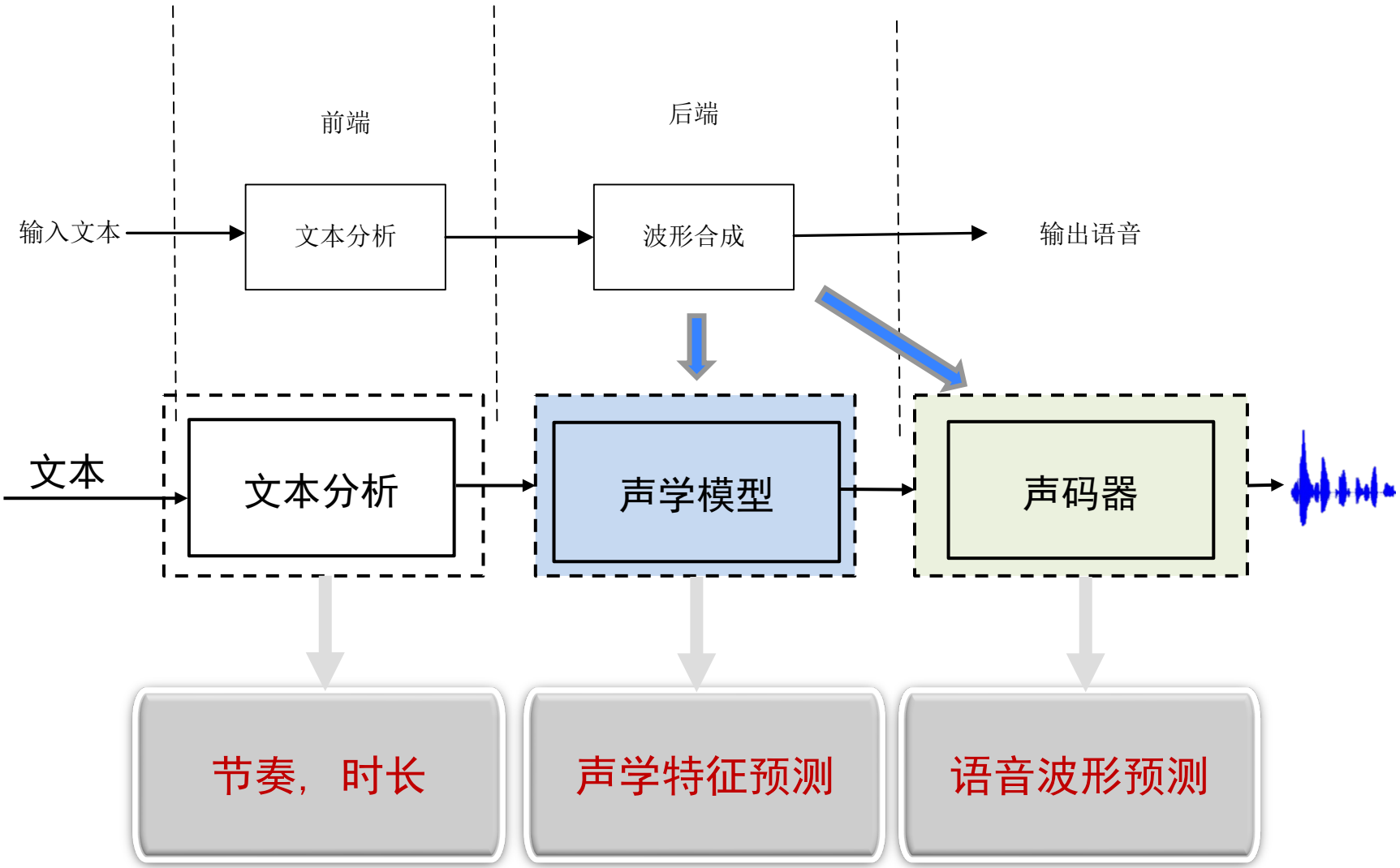
# 用于基元选取的上下文信息

- 需要提取每一个音节的上下文信息
  - 音节发音信息
  - 音节位置信息
  - 音节时长信息
  - 词位置信息
  - 韵律短语位置信息
  - 语调短语位置信息
  - 音节边界信息
  - 词性

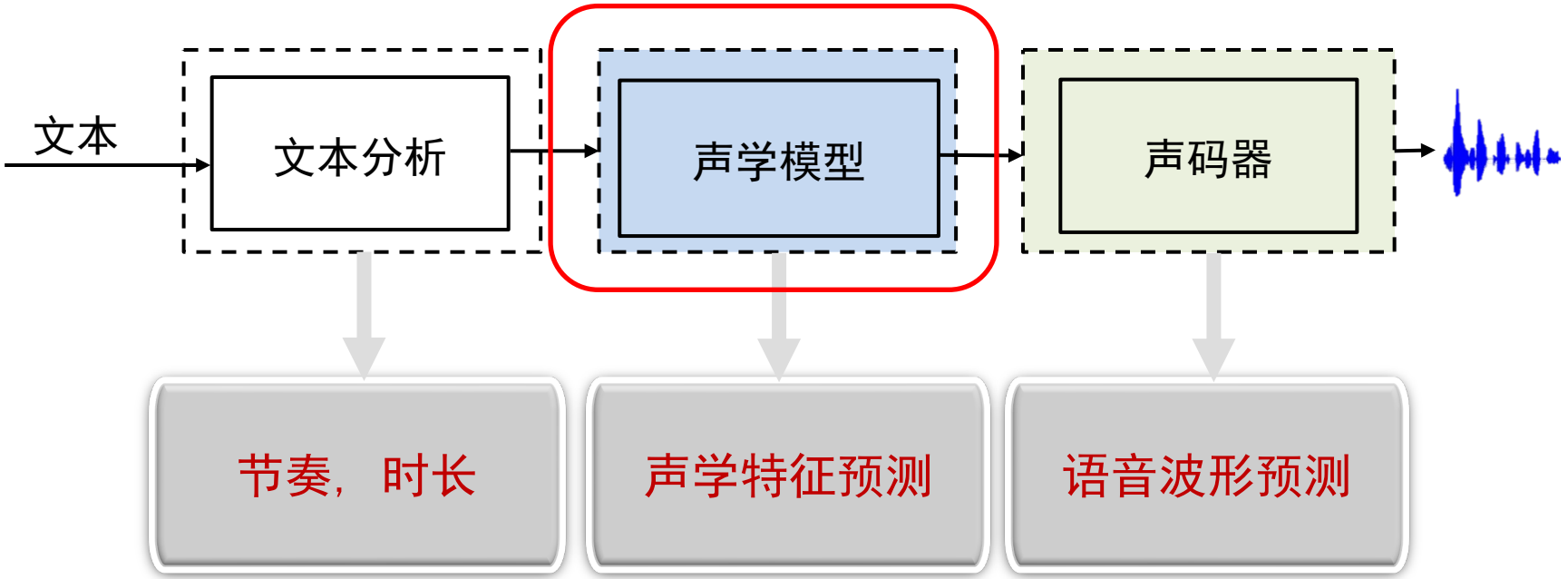
# 连续语句的基元选取方法



# 管道式语音合成



# 管道式语音合成



# 语音合成的声学特征预测

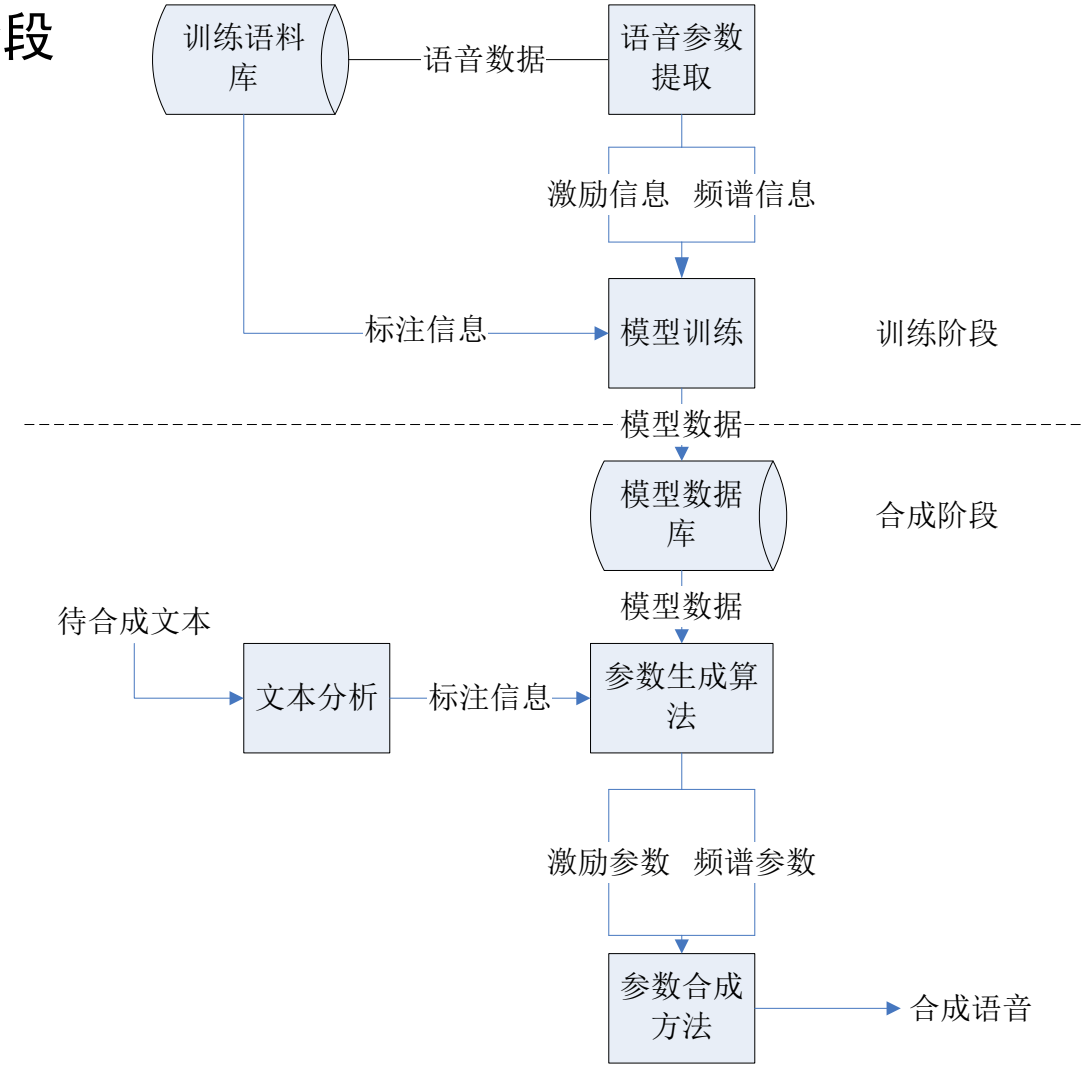
- 基于HMM的方法
- 基于深度神经网络的方法

# 语音合成的声学模型

- 基于HMM的方法
- 基于深度神经网络的方法

# 基于HMM的语音合成

- 基于HMM的语音合成方法主要分为两个阶段：
  - 训练阶段
  - 合成阶段

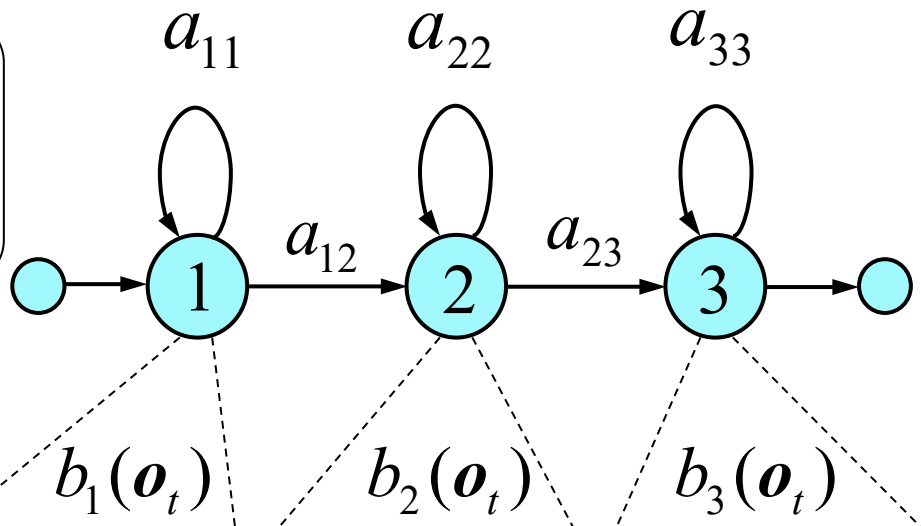


# 隐马尔可夫模型（HMM）

- 状态序列未知的马尔可夫链由观测估计状态序列

$a_{ij}$ ： 状态转移概率

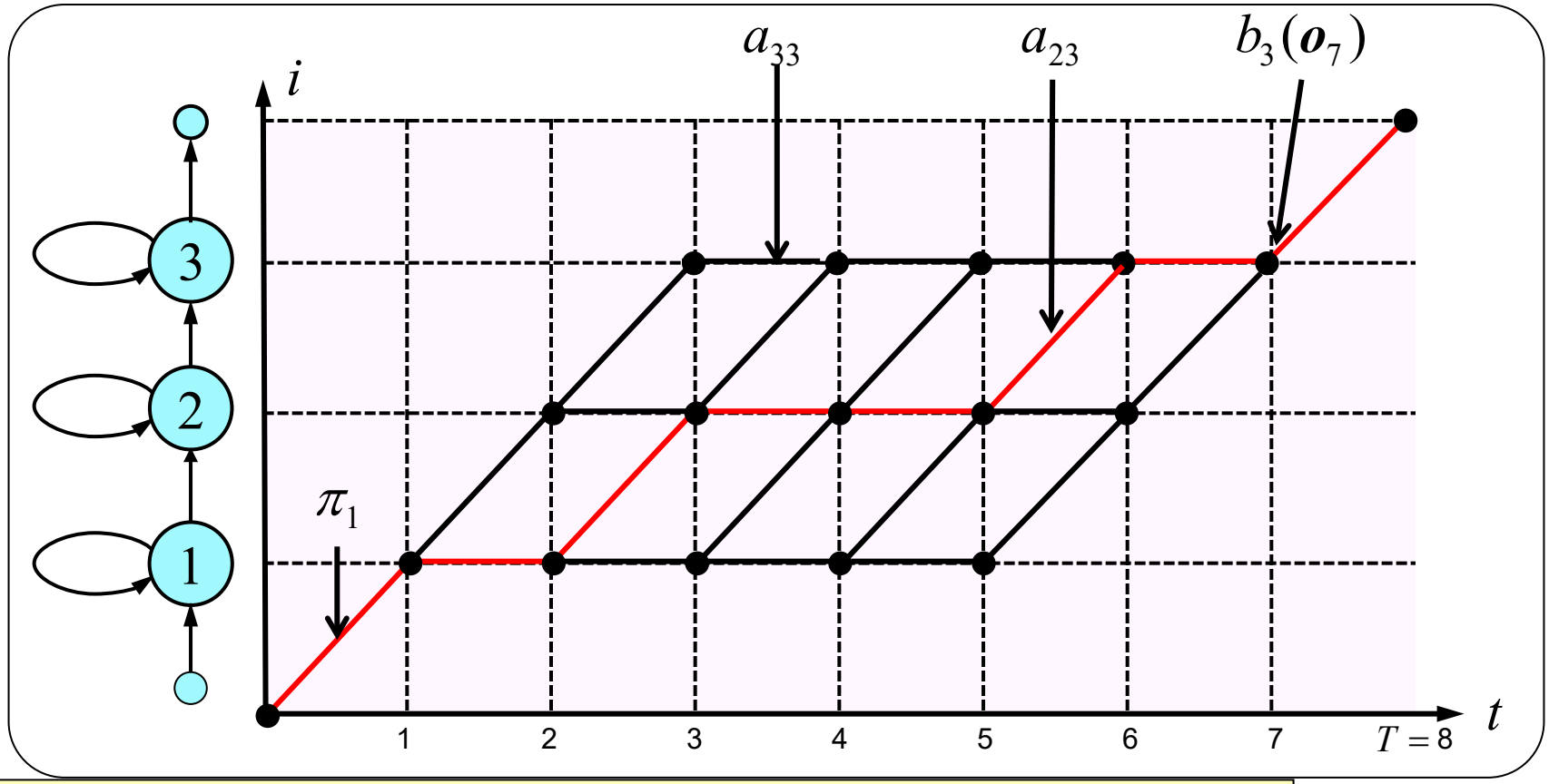
$b_q(\boldsymbol{o}_t)$ ： 观测概率




观测序列 $\boldsymbol{o}$ $\boldsymbol{o}_1$ $\boldsymbol{o}_2$ $\boldsymbol{o}_3$ $\boldsymbol{o}_4$ $\boldsymbol{o}_5$ … ┐ ┐ … $\boldsymbol{o}_T$

输出序列 $\boldsymbol{q}$ 1 1 1 1 2 … 2 3 … 3

# HMM 输出概率




似然函数

$$P(\boldsymbol{o}\mid\lambda)=\sum_{\boldsymbol{q}}P(\boldsymbol{o},\boldsymbol{q}\mid\lambda)=\sum_{\boldsymbol{q}}\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}(\boldsymbol{o}_t)$$

# 语音参数生成算法

- 对于给定的HMM $\lambda$, 确定一个语音参数向量
- 序列$\boldsymbol{o} = [\boldsymbol{o}_1^{\mathsf{T}}, \boldsymbol{o}_2^{\mathsf{T}}, \ldots, \boldsymbol{o}_T^{\mathsf{T}}]^{\mathsf{T}}$最大化

$$P(\boldsymbol{o} \mid \lambda) = \sum_{q} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

$$\approx \max_{q} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

$$\Downarrow$$

$$\hat{\boldsymbol{q}} = \arg\max_{q} P(\boldsymbol{q} \mid w, \lambda)$$

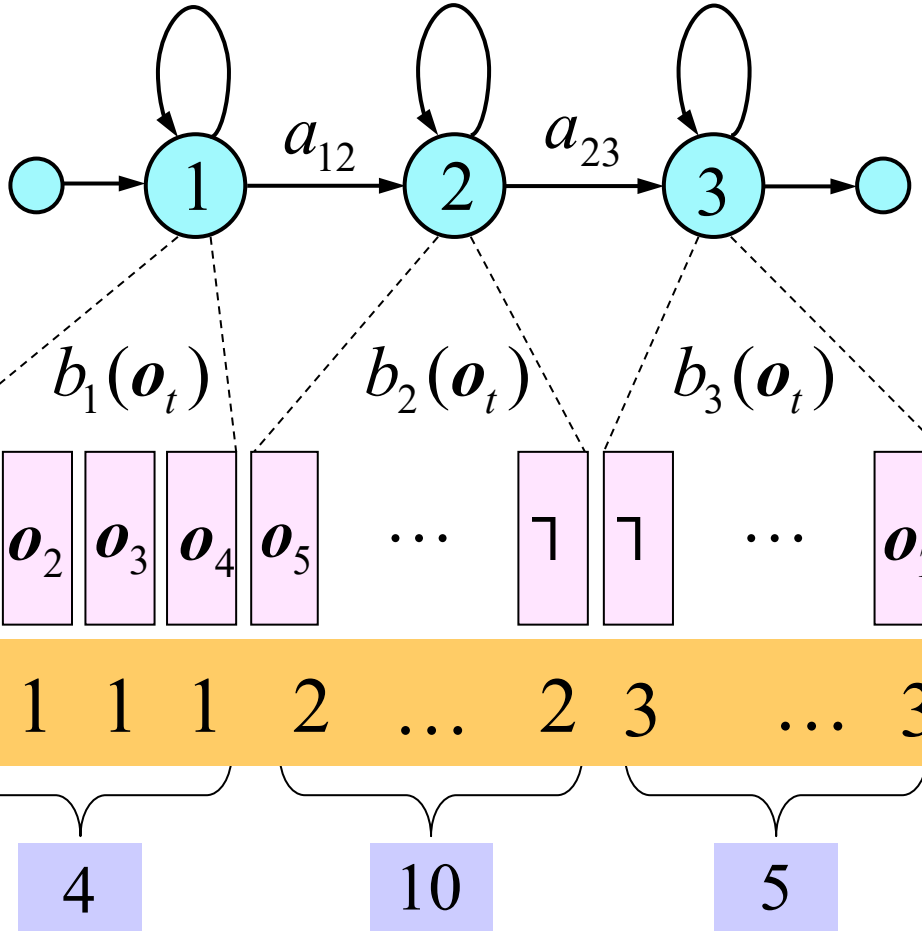
$$\hat{\boldsymbol{o}} = \arg\max_{o} P(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \lambda)$$

# 确定状态序列

- 通过确定状态持续时间来确定状态序列

$a_{ij}$ ： 状态转移概率

$b_q(\boldsymbol{o}_t)$ ： 输出概率



# 确定状态序列

$$P(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}}) = \prod_{i=1}^{K} p_i(d_i)$$

$p_i(\cdot)$ ：$i$ –th 状态持续时间分布

$d_i$ ： –th 状态持续时间

$K$ ：对于 $w$ 的HMM句子状态

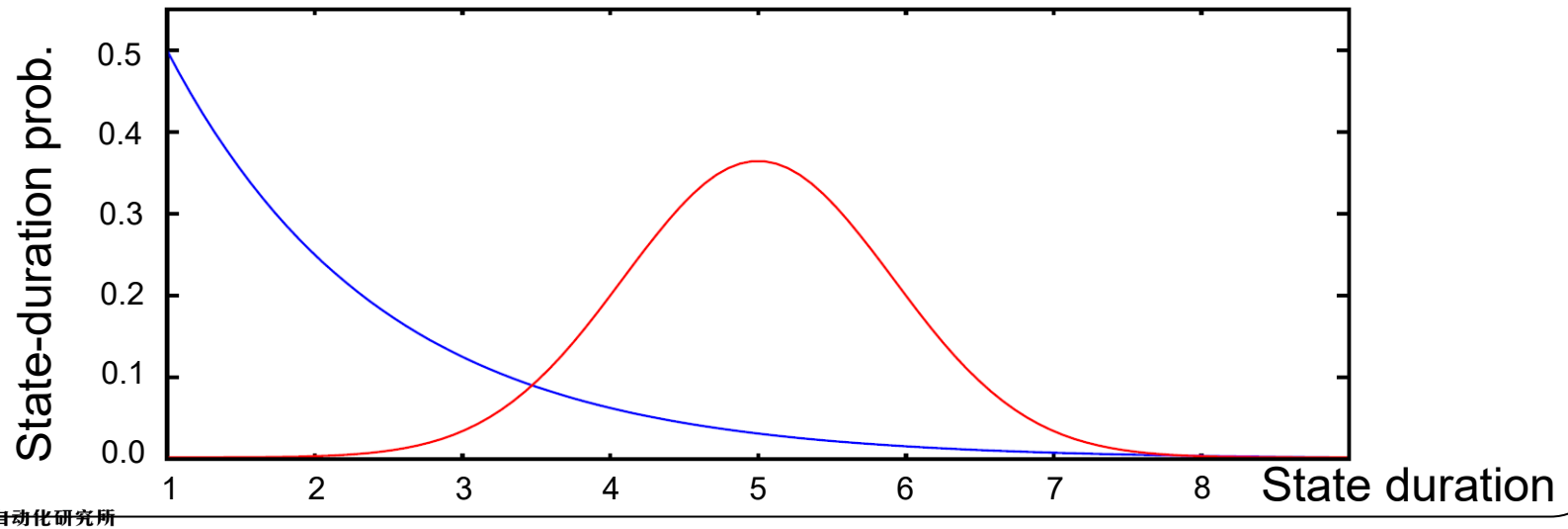
# 确定状态序列

## Geometric

$$p_i(d_i) = a_{ij}^{d_i - 1}(1 - a_{ii}) \quad \Rightarrow \quad \hat{d}_i = 1$$

## Gaussian

$$p_i(d_i) = N(d_i \mid m_i, \sigma_i^2) \quad \Rightarrow \quad \hat{d}_i = m_i$$



# 语音参数生成算法

- 对于给定的HMM $\lambda$, 确定一个语音参数向量
- 序列$\boldsymbol{o} = [\boldsymbol{o}_1^{\mathsf{T}}, \boldsymbol{o}_2^{\mathsf{T}}, \ldots, \boldsymbol{o}_T^{\mathsf{T}}]^{\mathsf{T}}$最大化

$$P(\boldsymbol{o} \mid \lambda) = \sum_{q} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

$$\approx \max_{q} P(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) P(\boldsymbol{q} \mid \lambda)$$

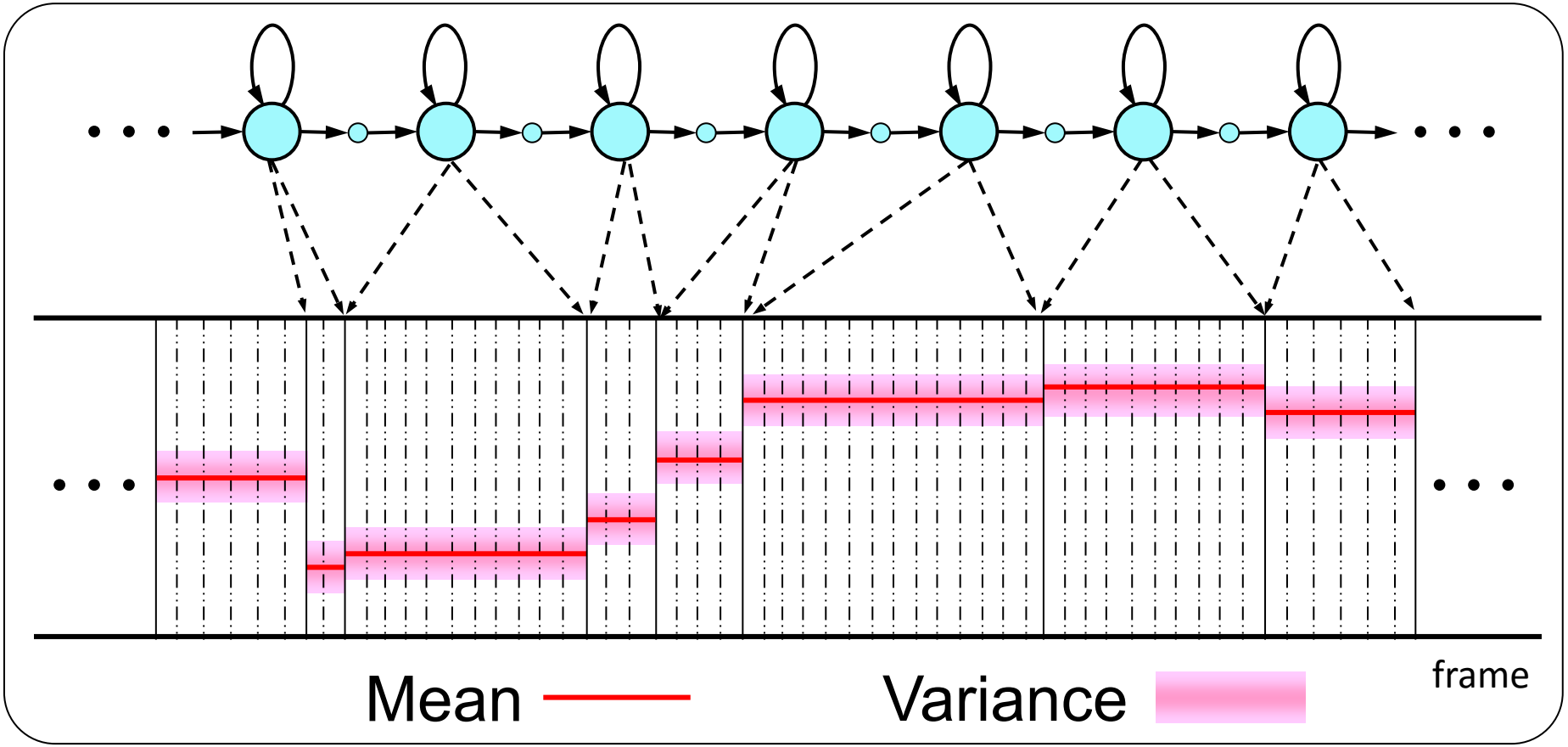
$$\Downarrow$$

$$\hat{\boldsymbol{q}} = \arg\max_{q} P(\boldsymbol{q} \mid w, \lambda)$$

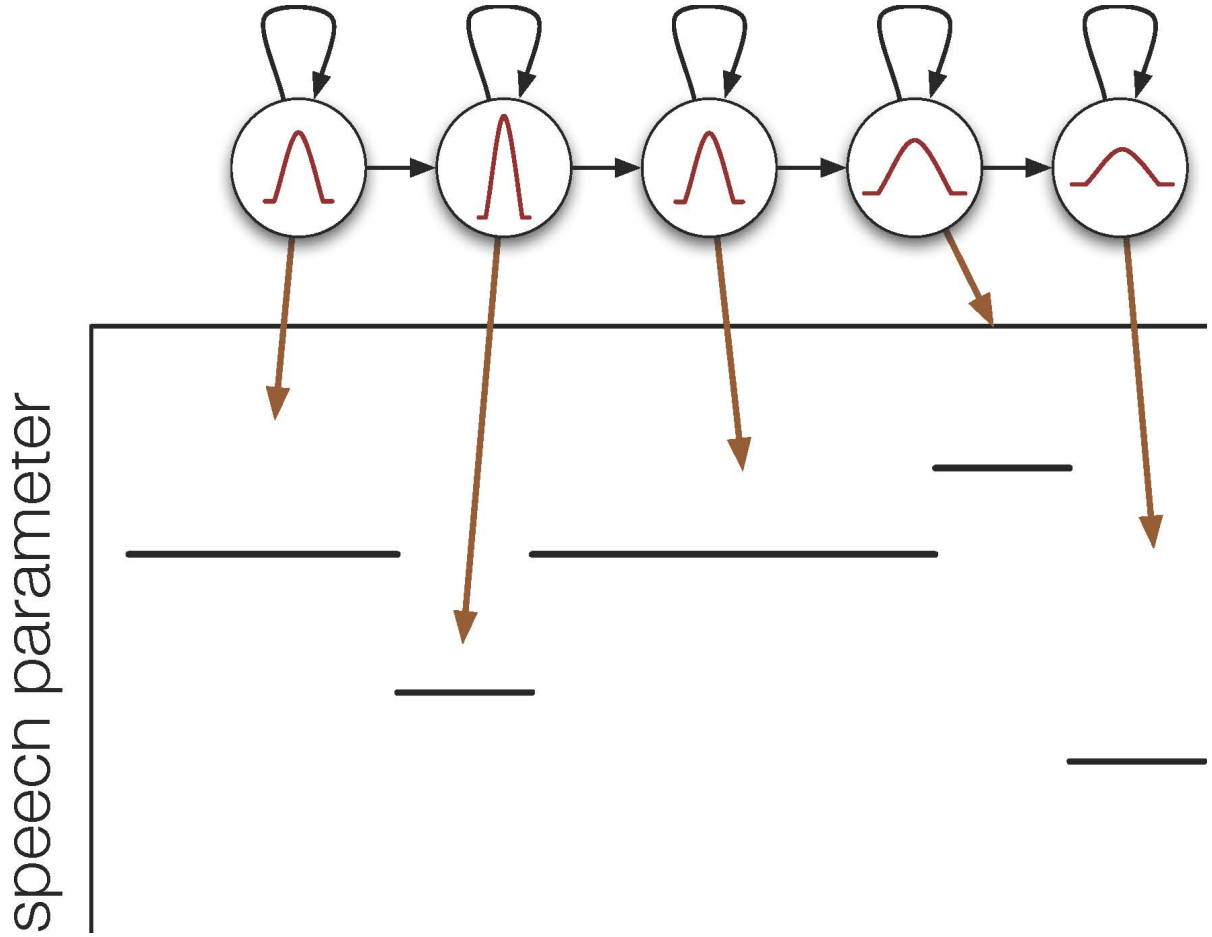
$$\hat{\boldsymbol{o}} = \arg\max_{o} P(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \lambda)$$

# 生成特征序列

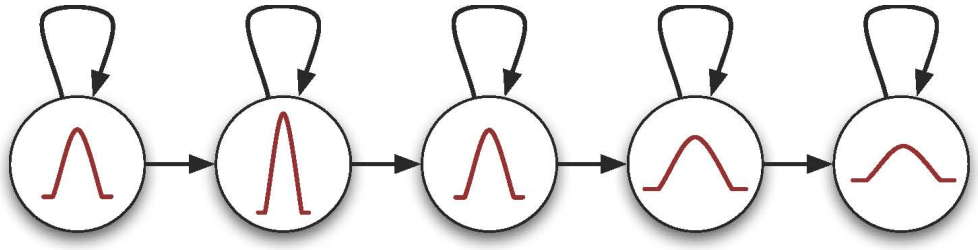
- $\hat{\mathbf{o}}$ 变成一个均值向量序列
- 显示状态之间的非连续输出

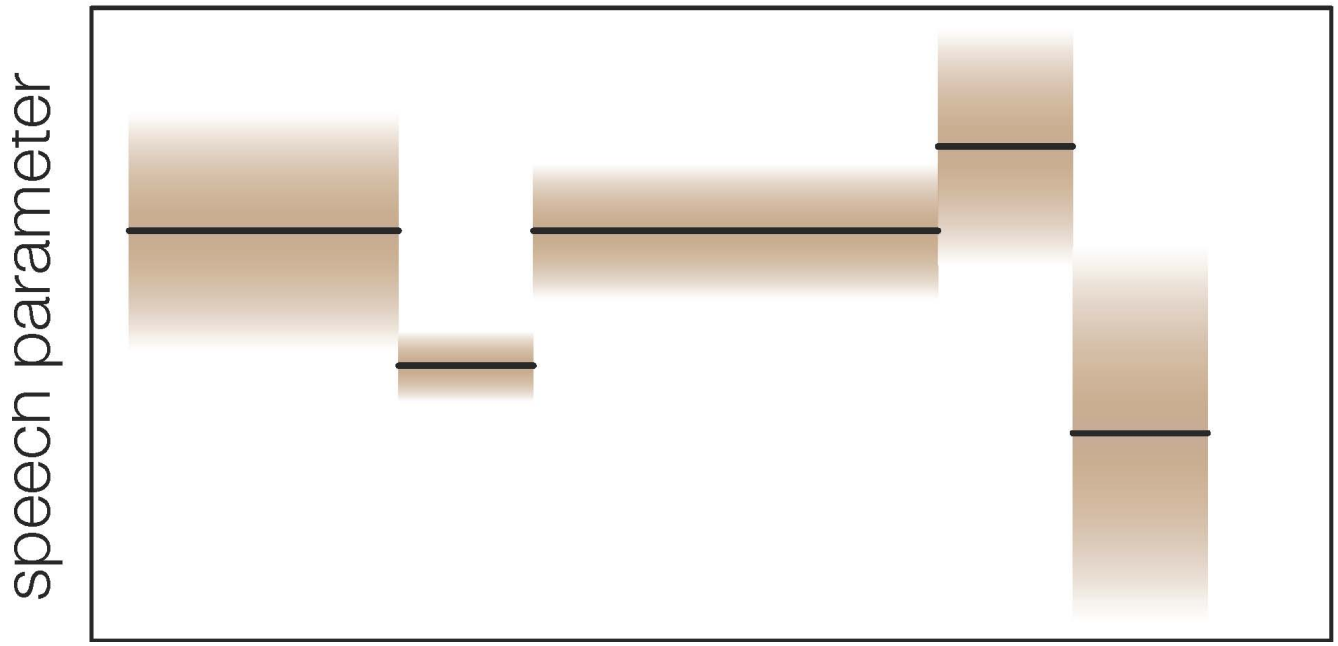


# HMM参数生成

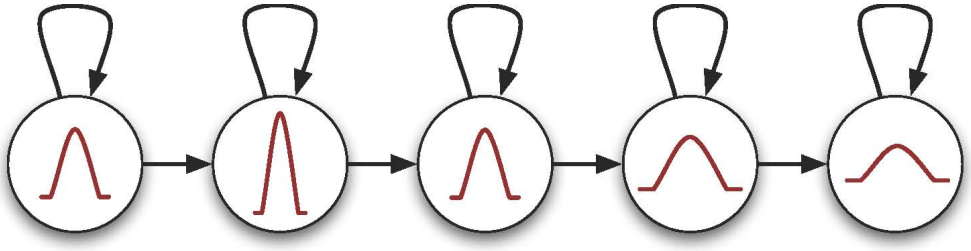


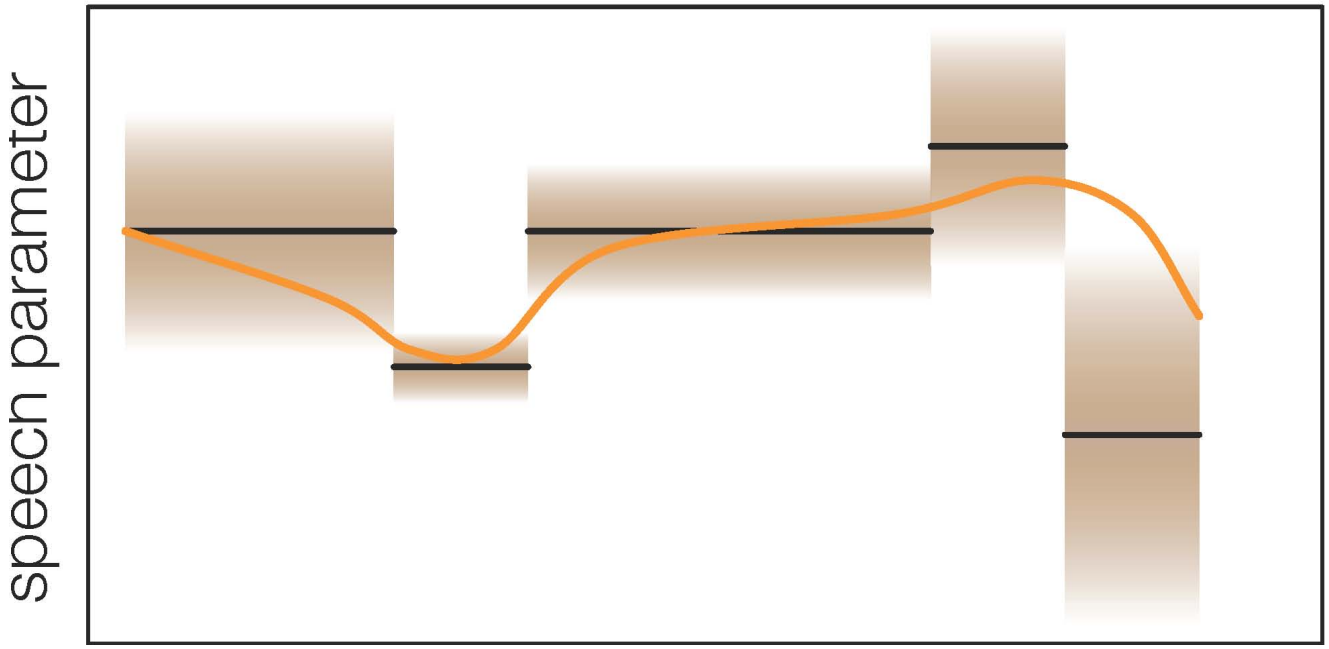
# HMM参数生成





# HMM参数生成

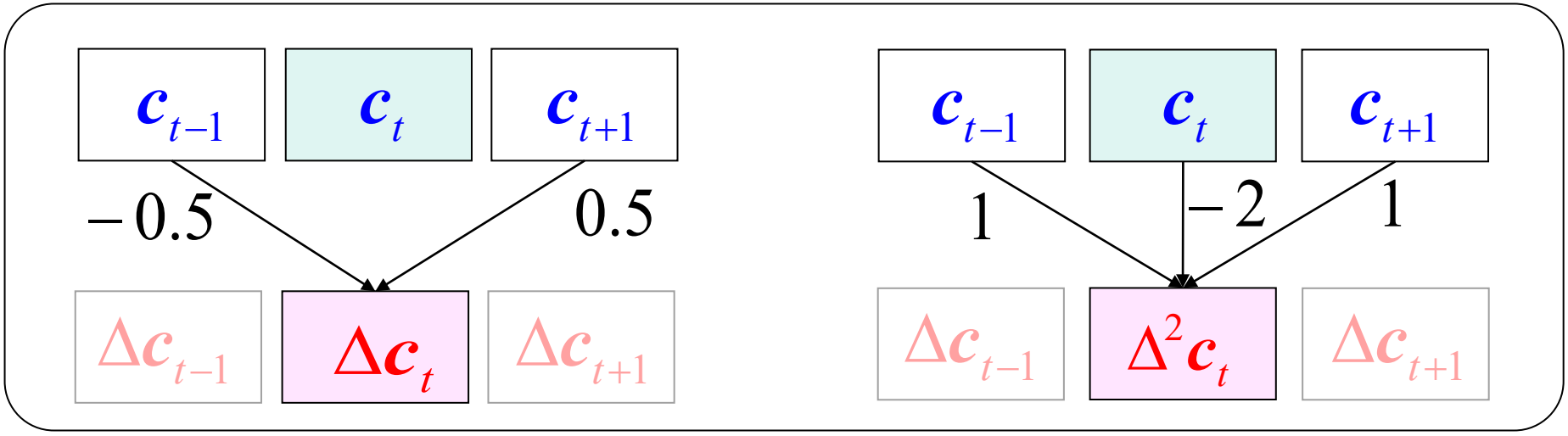




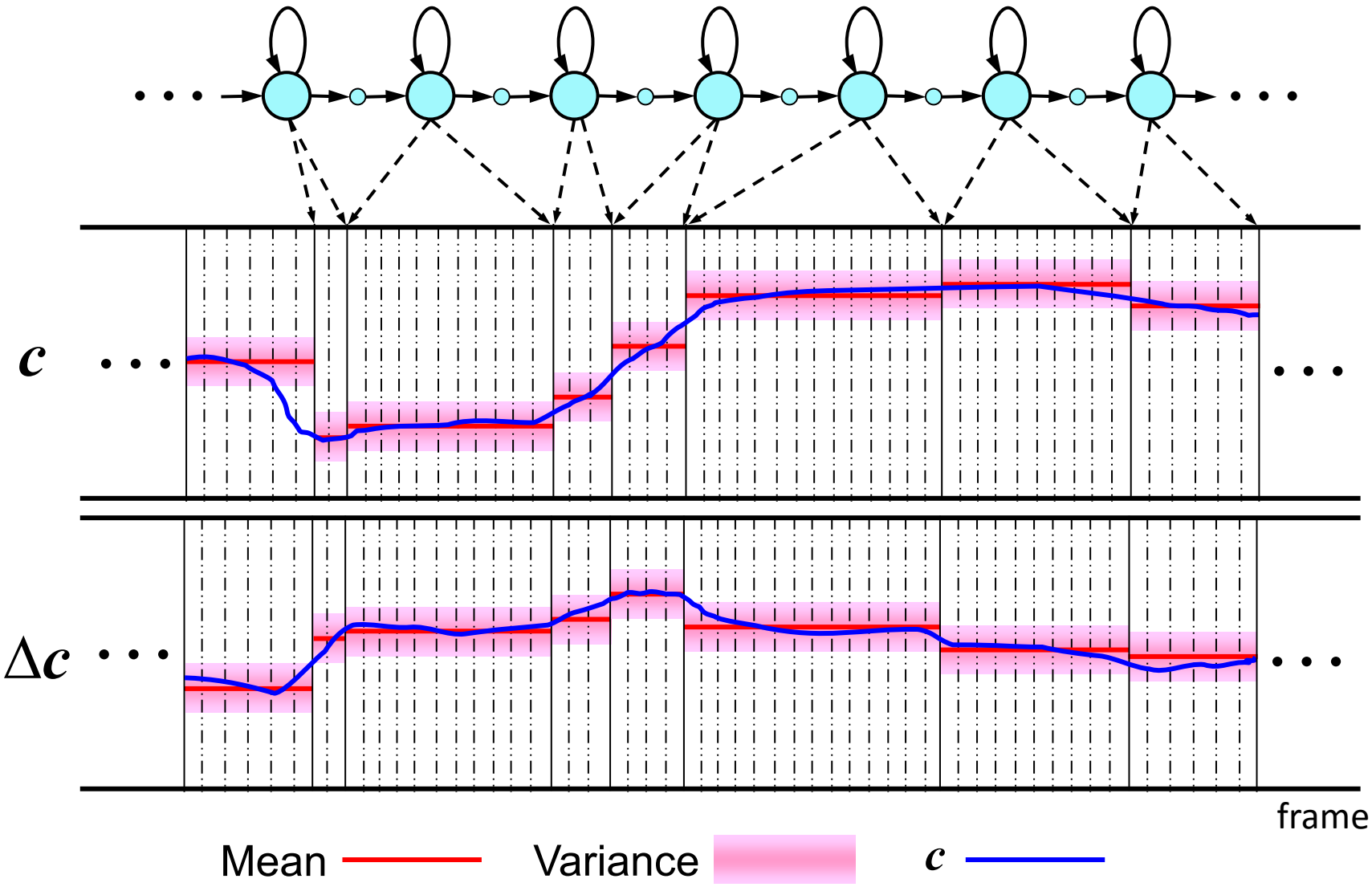
# 动态特征

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1})$$

$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1}$$

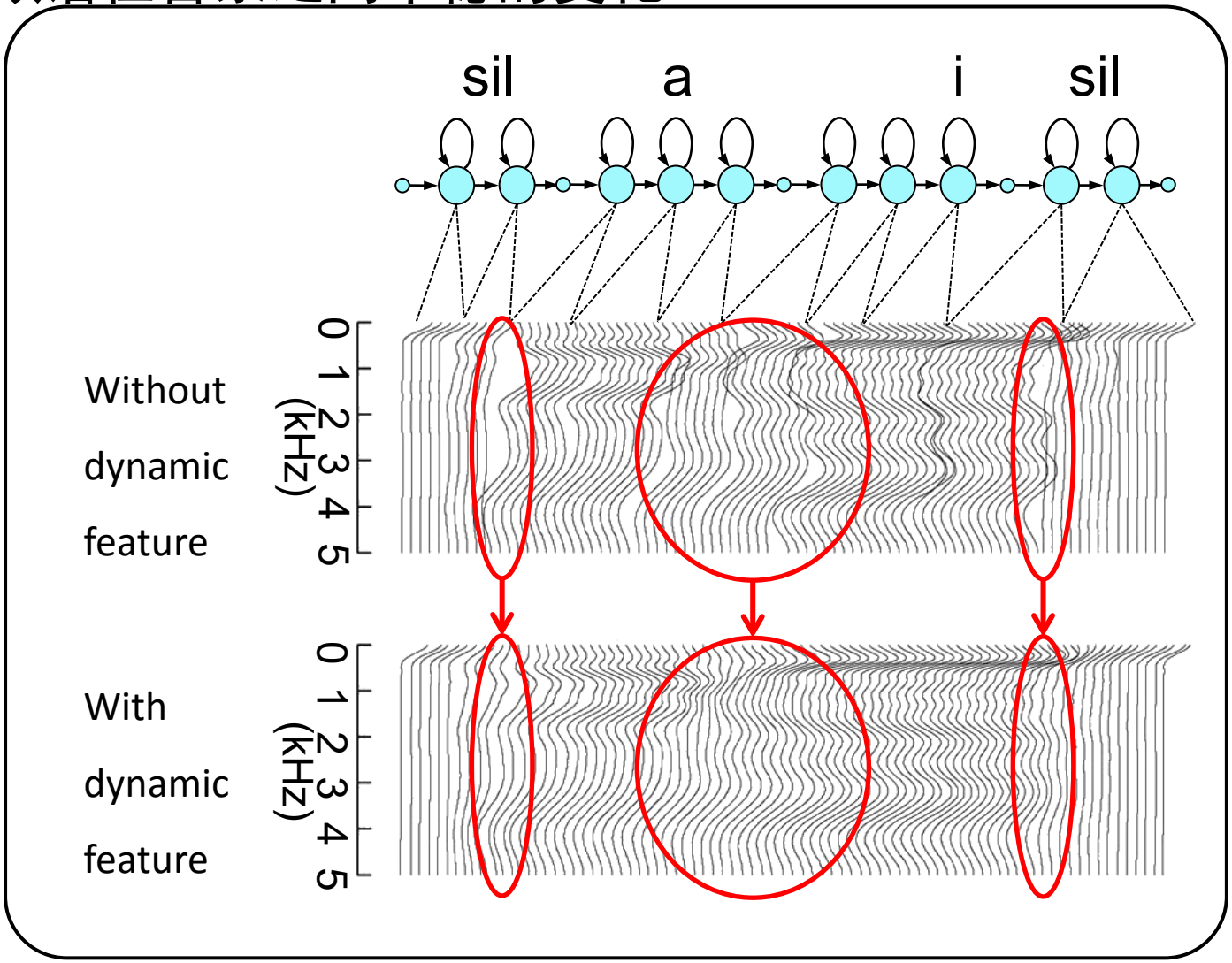


# 生成语音参数轨迹



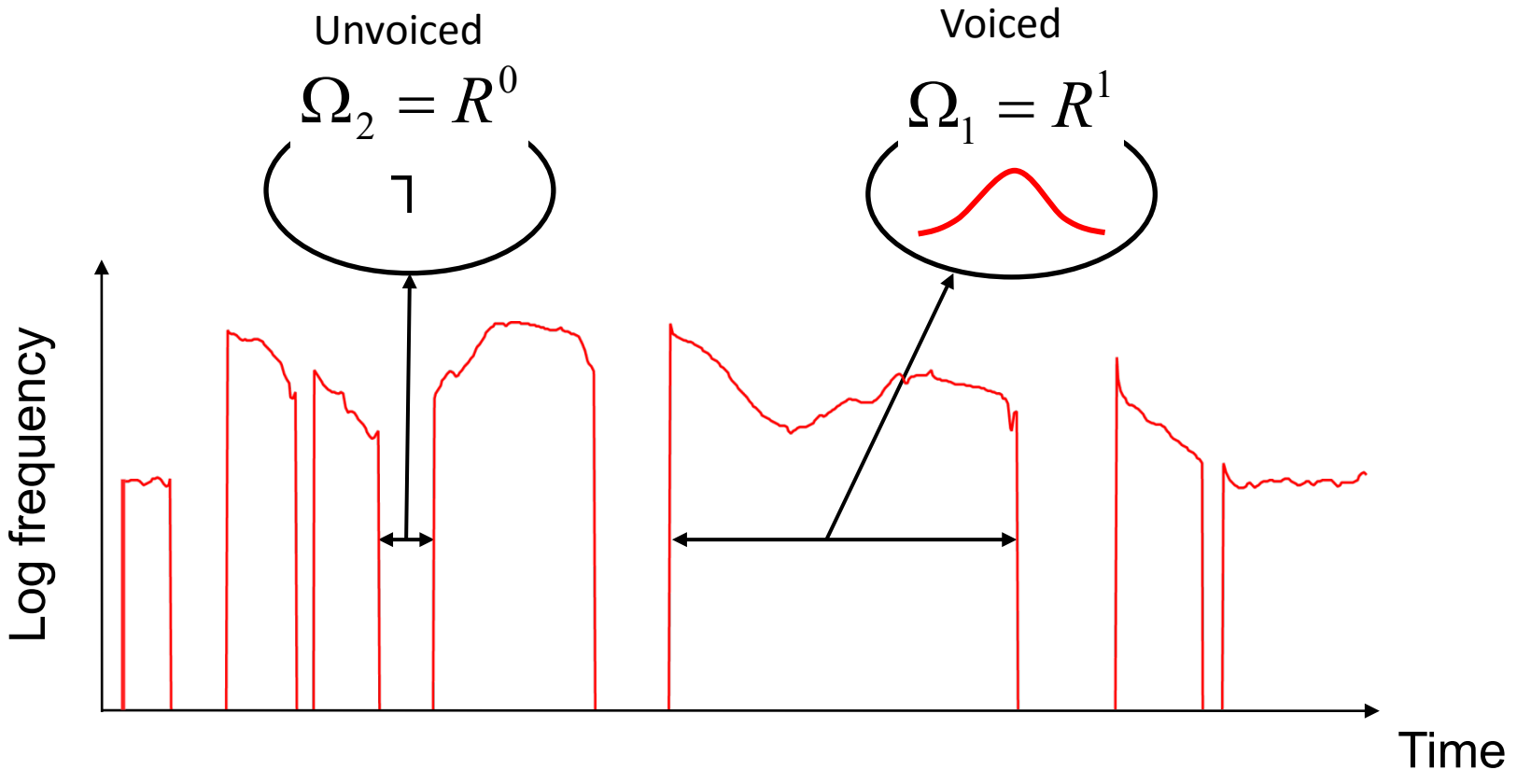
# 生成频谱图

- 频谱在音素之间平稳的变化



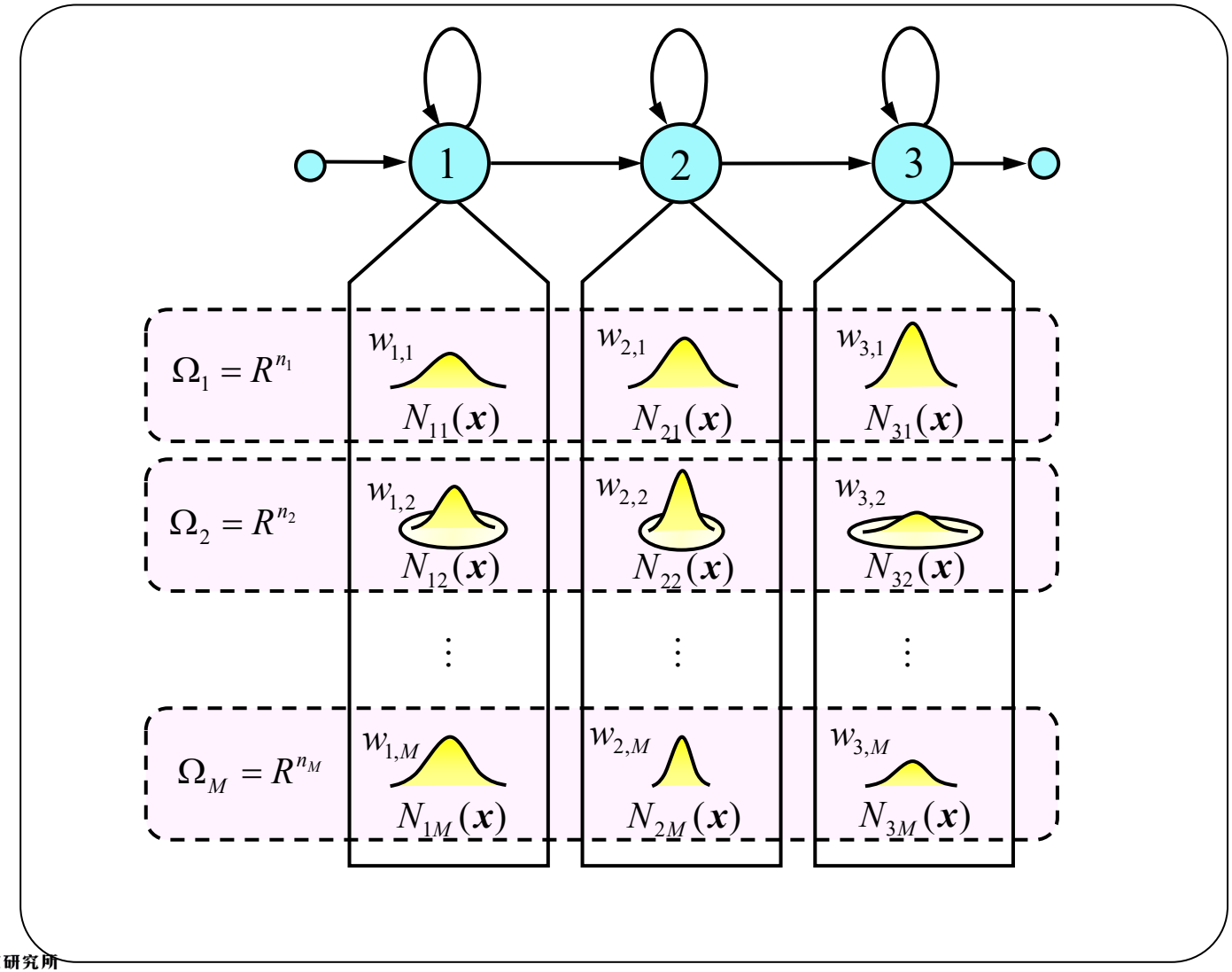
# F0观测

- 无法通过连续或离散分布进行建模
- 多空间概率分布HMM （MSD-HMM）



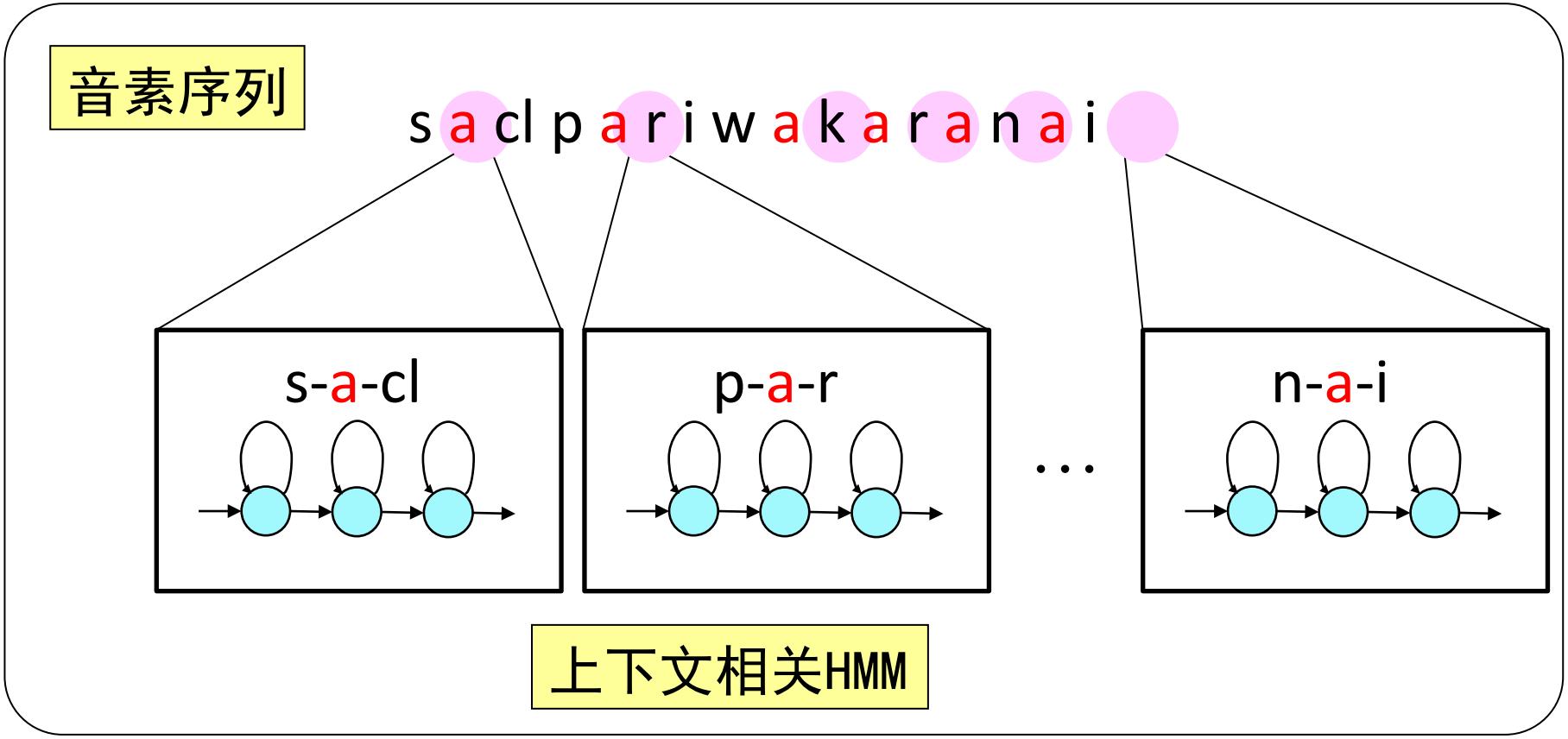
# MSD-HMM结构

- 每个状态有两个或更多的分布



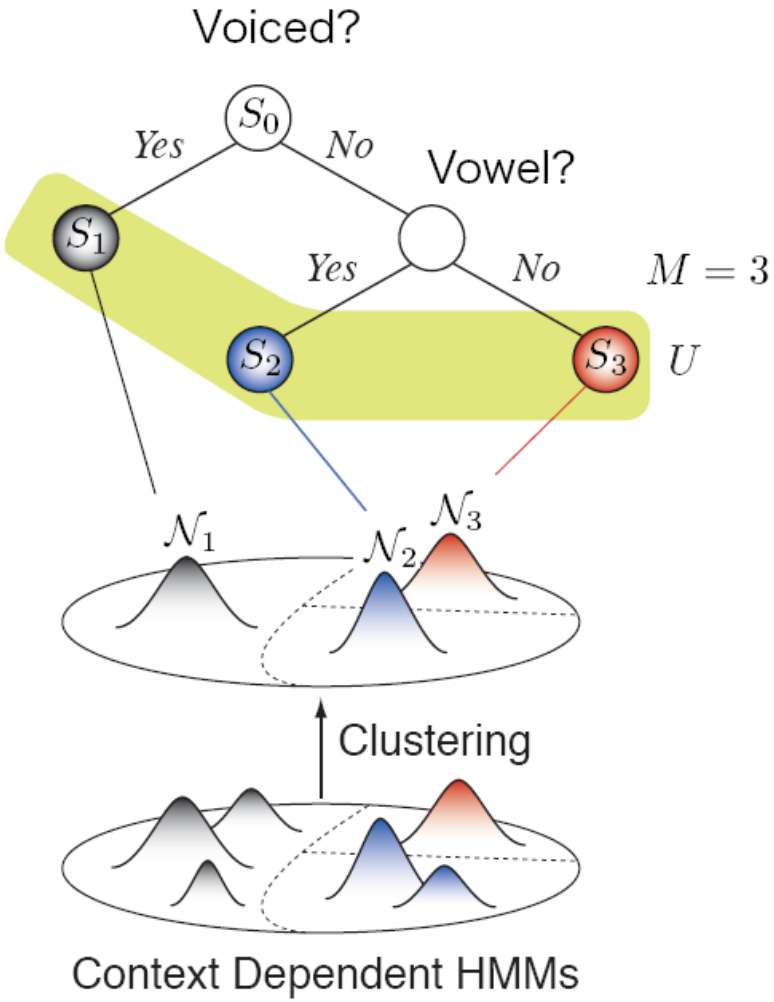
# 上下文相关建模

- 考虑音素之间的关系



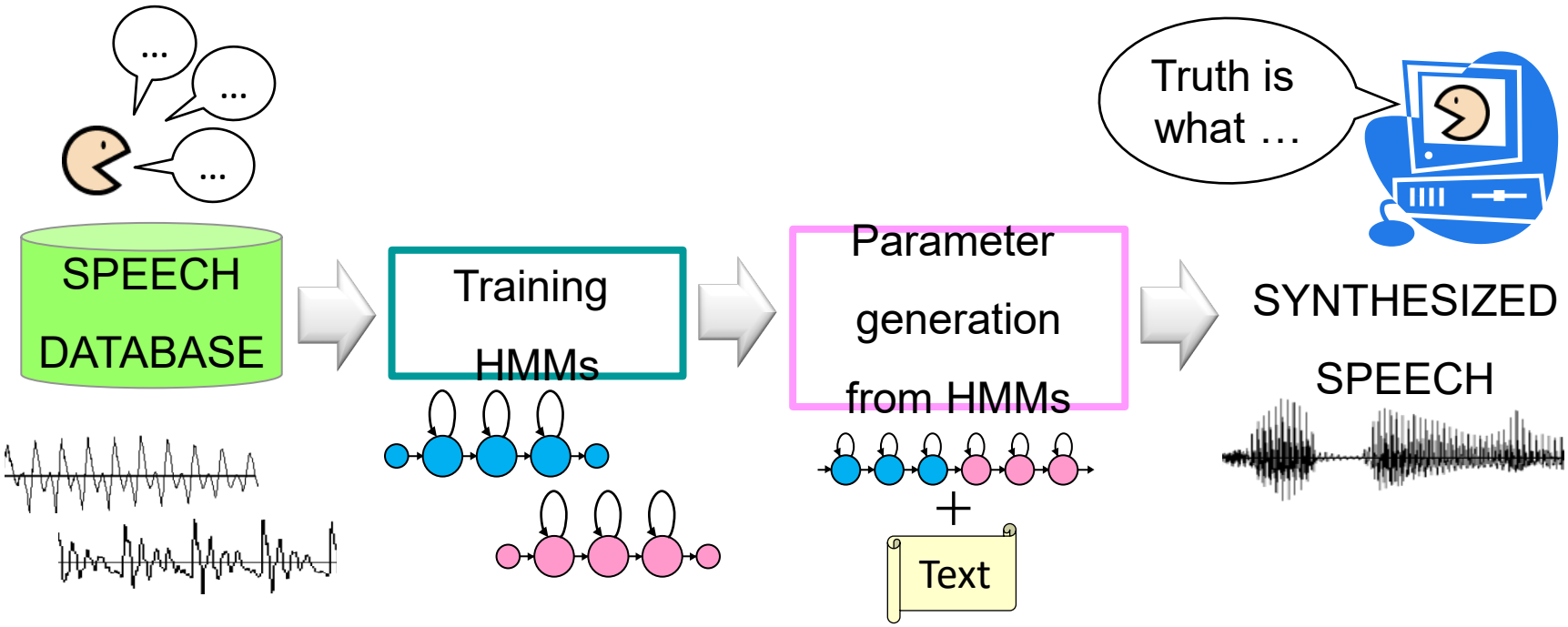
# 上下文相关建模

- 利用决策树对上下文相关的HMM进行聚类



# HTS步骤

- 声学特征由HMMs生成
  - 频谱参数
  - 基频参数（F0）
- 声码器合成语音
  - 获得高音质的合成语音



# 语音合成的声学模型

- 基于HMM的方法
- 基于深度神经网络的方法

# 基于深度学习的声学模型

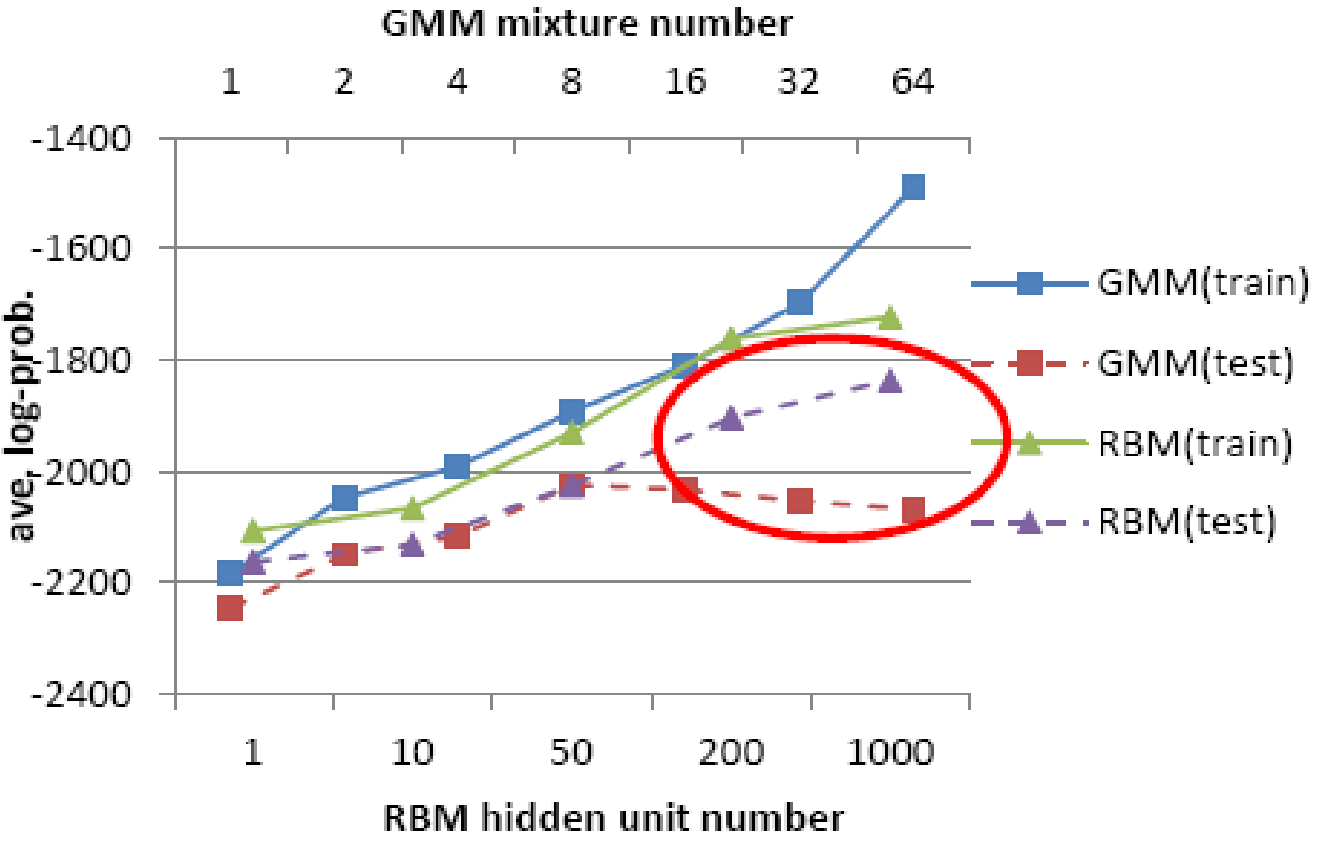
## 分类

- 受限玻尔兹曼机模型（RBM）
- 深度前馈网络（DNN）
- 深度混合密度网络（DMDN）
- 深度循环神经网络（RNN）

# 基于RBM的声学模型

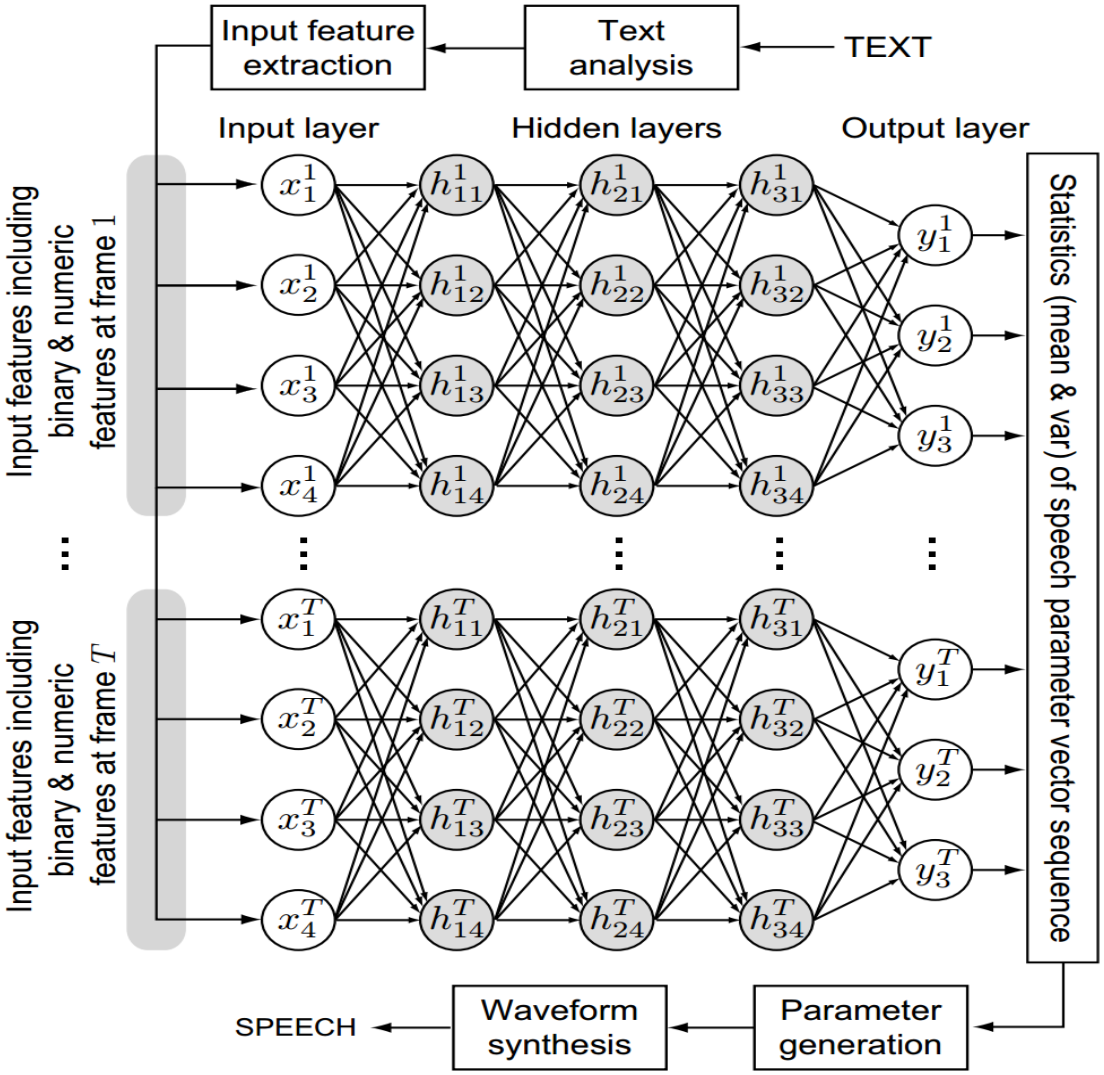
- GMM $p(v) = \sum_{i=1}^{M} m_i N(u_i, \sigma_i)$
  - 高斯混合密度模型产生的分布不会比组成它的单高斯成分更“尖锐”（过平滑）
  - 需要很多的数据来估计模型参数
- RBM $p(\mathbf{v}) \propto e^{-\frac{1}{2}(\mathbf{v}-\mathbf{b})^{\mathrm{T}}(\mathbf{v}-\mathbf{b})} \prod_{j} \left(1 + e^{c_j + \mathbf{v}^{\mathrm{T}}\mathbf{W}_{*,\mathrm{j}}}\right)$
  - 可以产生比单个成分更“尖锐”的分布
  - 只需要少量数据就可以估计模型参数

# 基于RBM的声学模型




spectral envelopes

# 基于DNN的声学模型


Input feature extraction
Text analysis
TEXT
Input layer
Hidden layers
Output layer
Input features including binary & numeric features at frame 1
Input features including binary & numeric features at frame T
Statistics (mean & var) of speech parameter vector sequence
SPEECH
Waveform synthesis
Parameter generation

# 基于DNN的声学模型

## 利用DNN构建文本特征到语音特征之间的映射

不管以前谁
对谁错

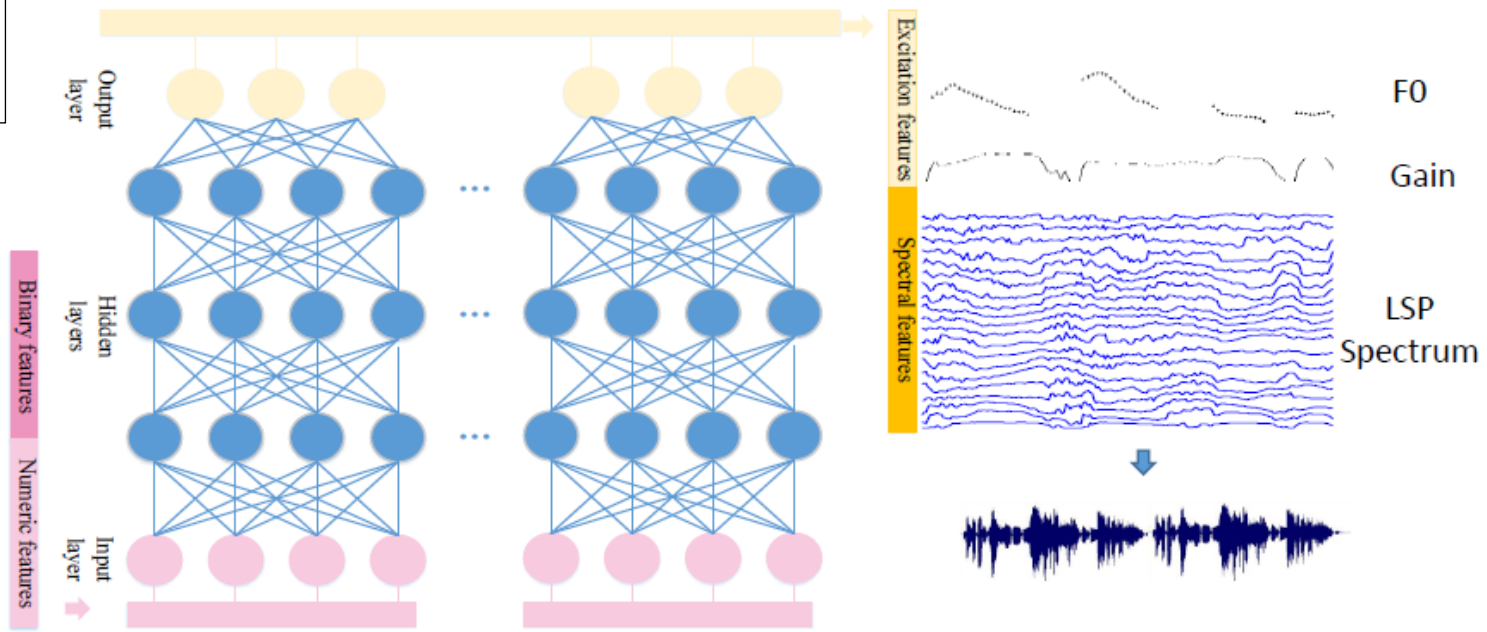
音素标签

词性标签

...

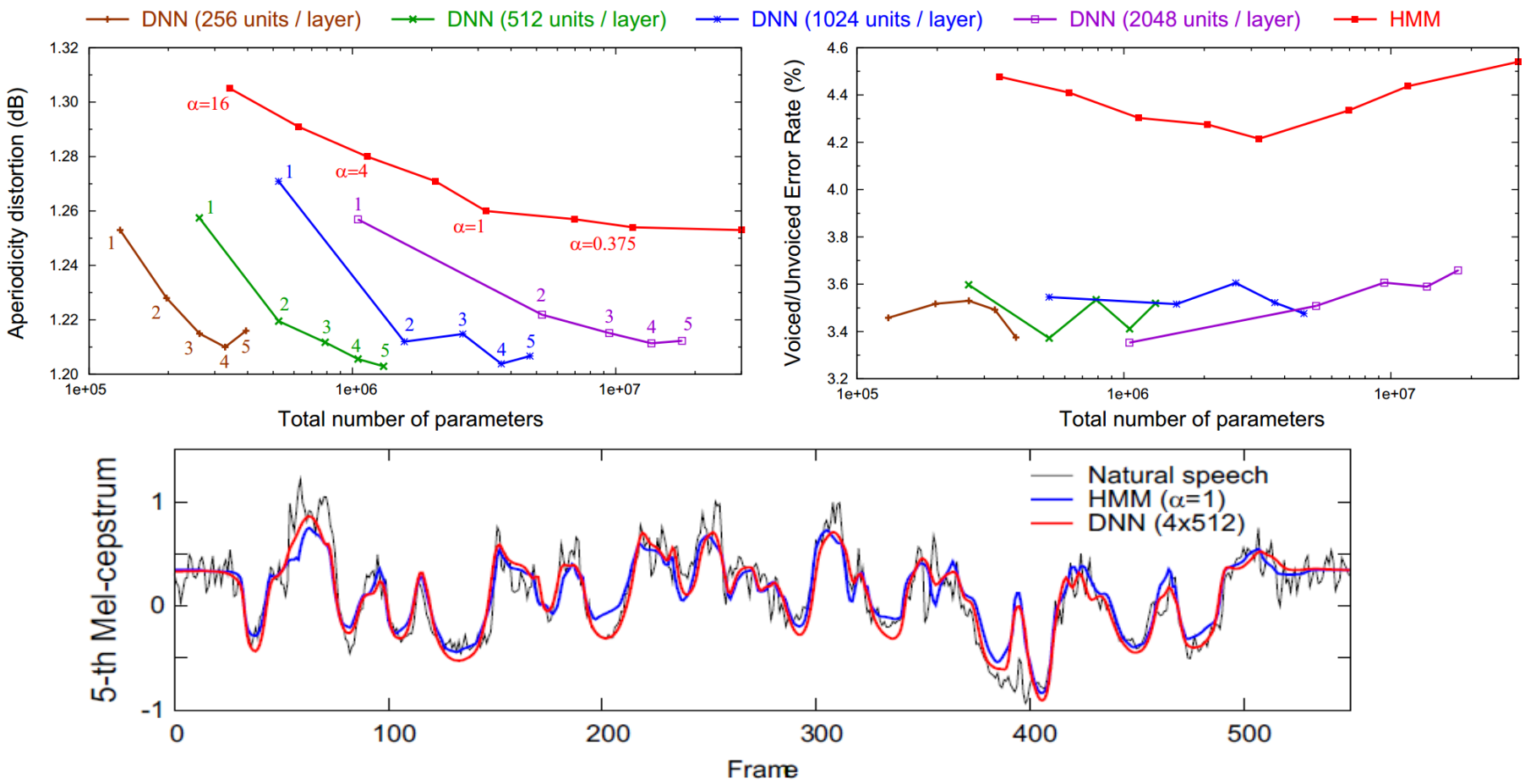
词数

时长




- 预测静态及一二阶语音参数
- 方差是通过设定为0.01倍的全局方差得到

# 基于DNN的声学模型



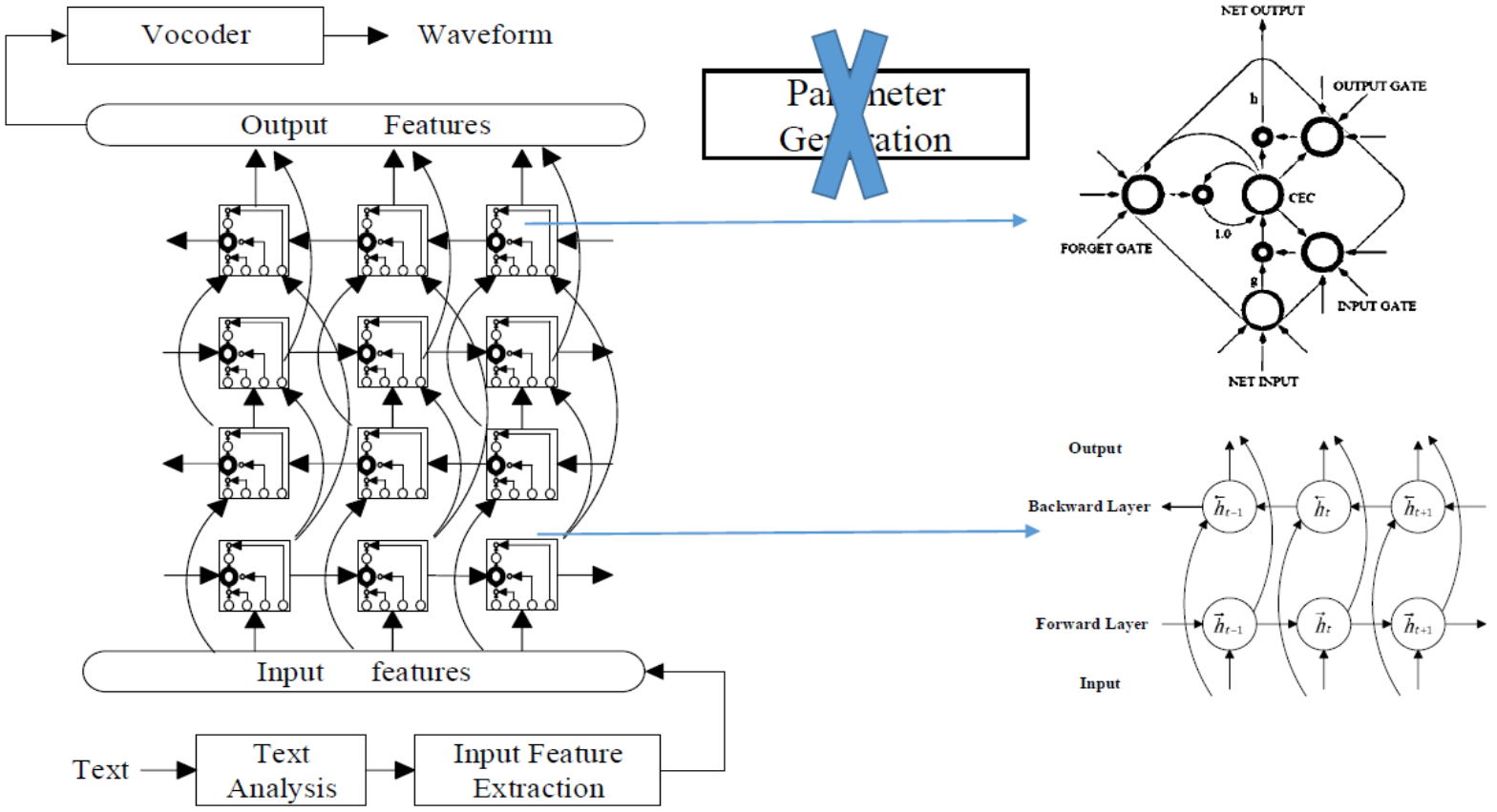

Ze, Heiga, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks." ***2013 ieee international conference on acoustics, speech and signal processing***. IEEE, 2013.

# 基于LSTM-RNN的声学模型

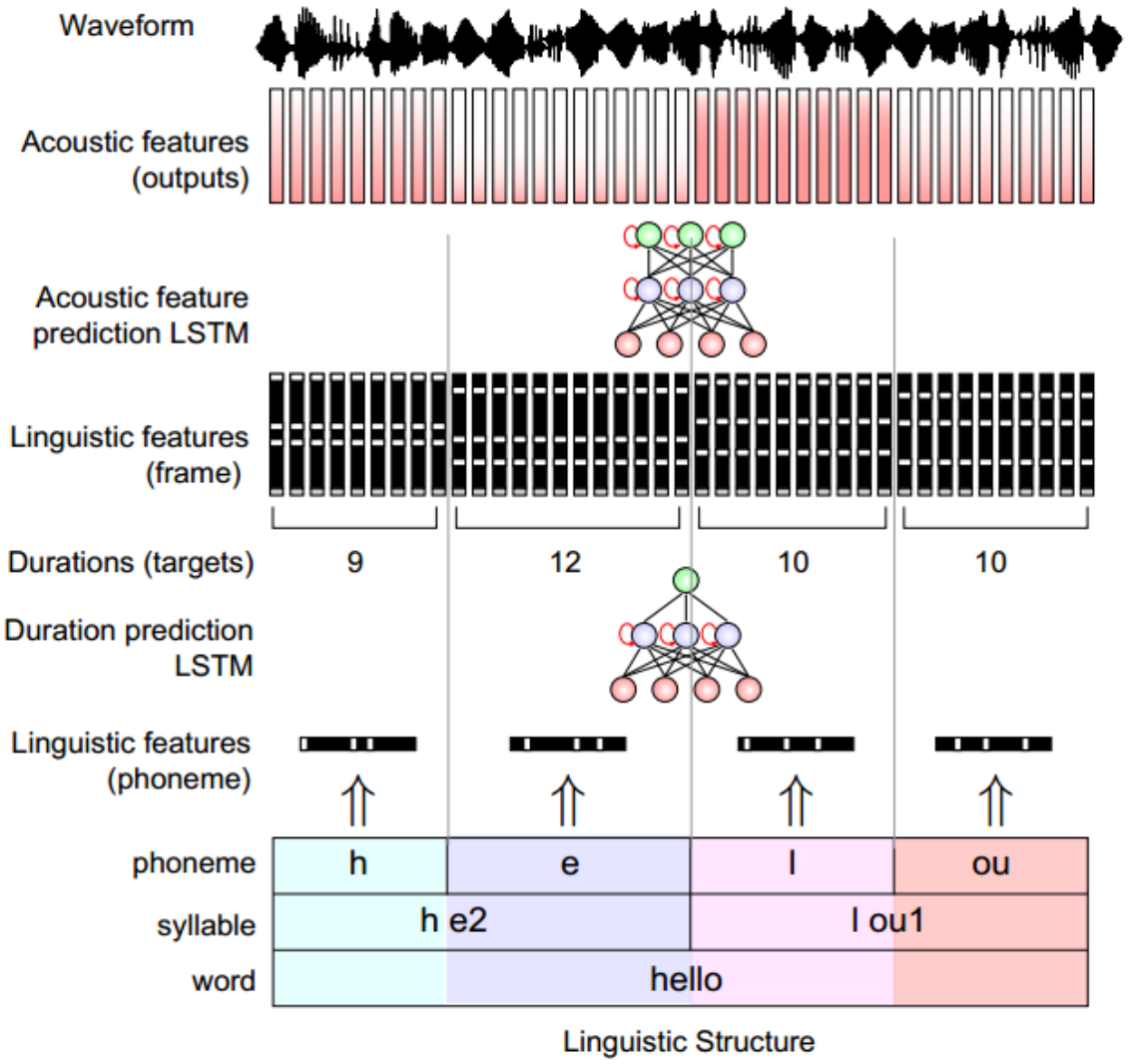
- 语音的生成是一个连续的动态过程
  - 考虑了语意、句法、词性等信息
  - 这些信息与其所在的上下文信息关联性很强
- LSTM-RNN
  - 能够综合考虑过去和未来的信息(DNN只能考虑固定的窗口长度)

# 基于LSTM-RNN的声学模型

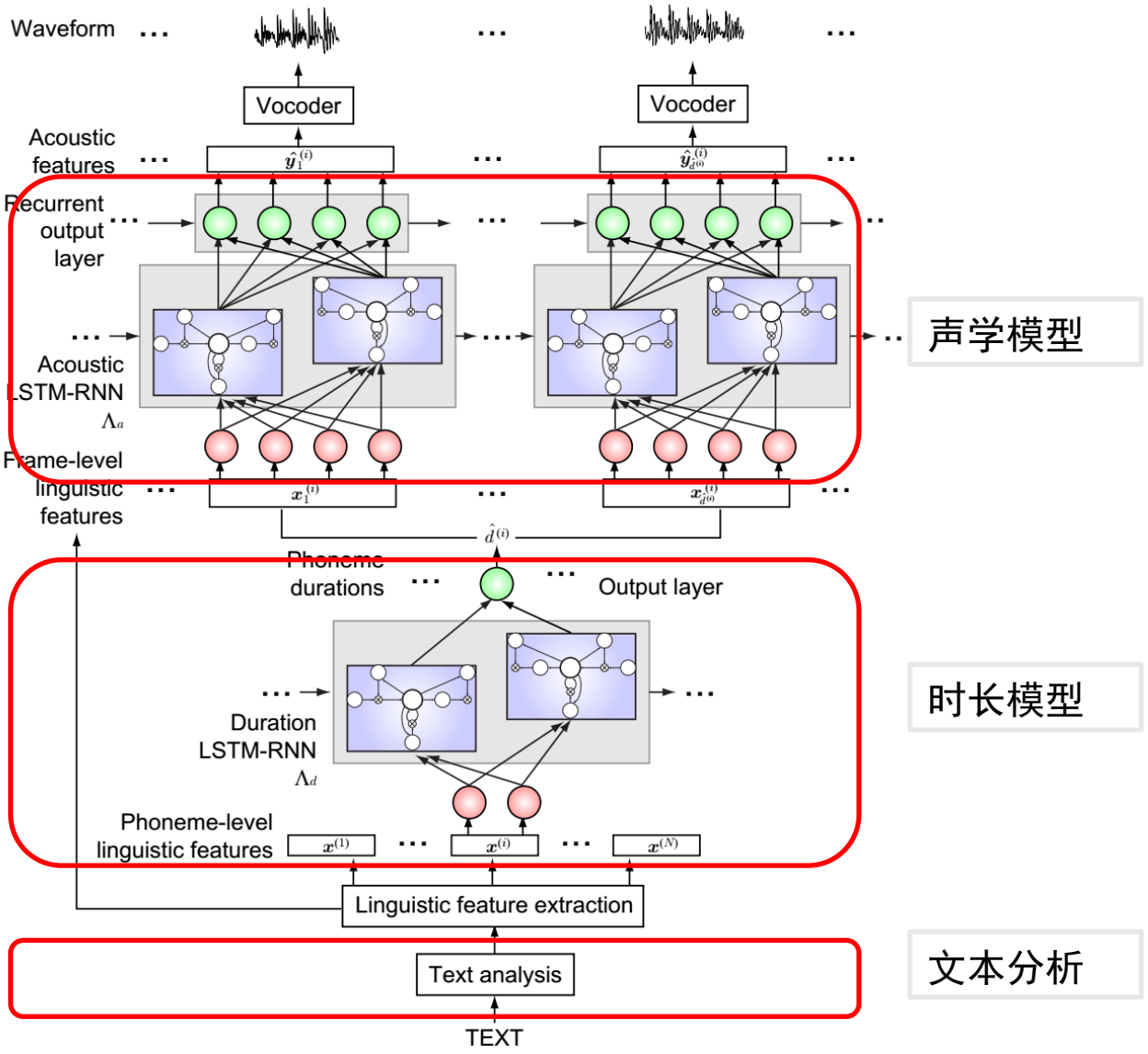
- 之前DNN声学模型不包括参数生成算法



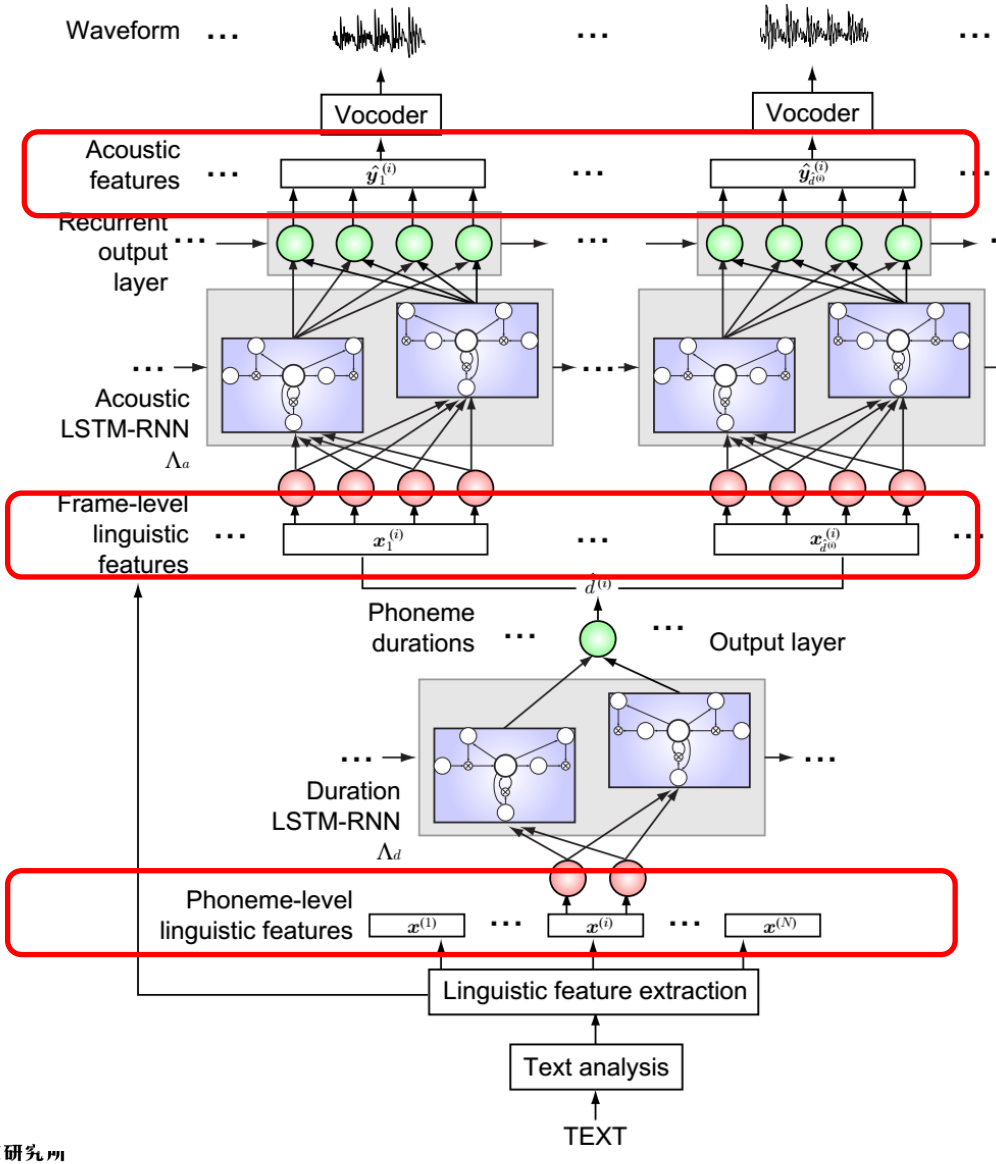
# 基于LSTM-RNN的声学模型



# 基于LSTM-RNN的声学模型



# 基于LSTM-RNN的声学模型




**声学参数：**

40维梅尔谱系数、基频等

**帧级别文本特征：**

在音素级文本特征的基础上，增加帧的位置信息，一个音素包含多帧

**音素级别文本特征：**

音素类型、重音位置、音节个数、音节在短语中的位置等

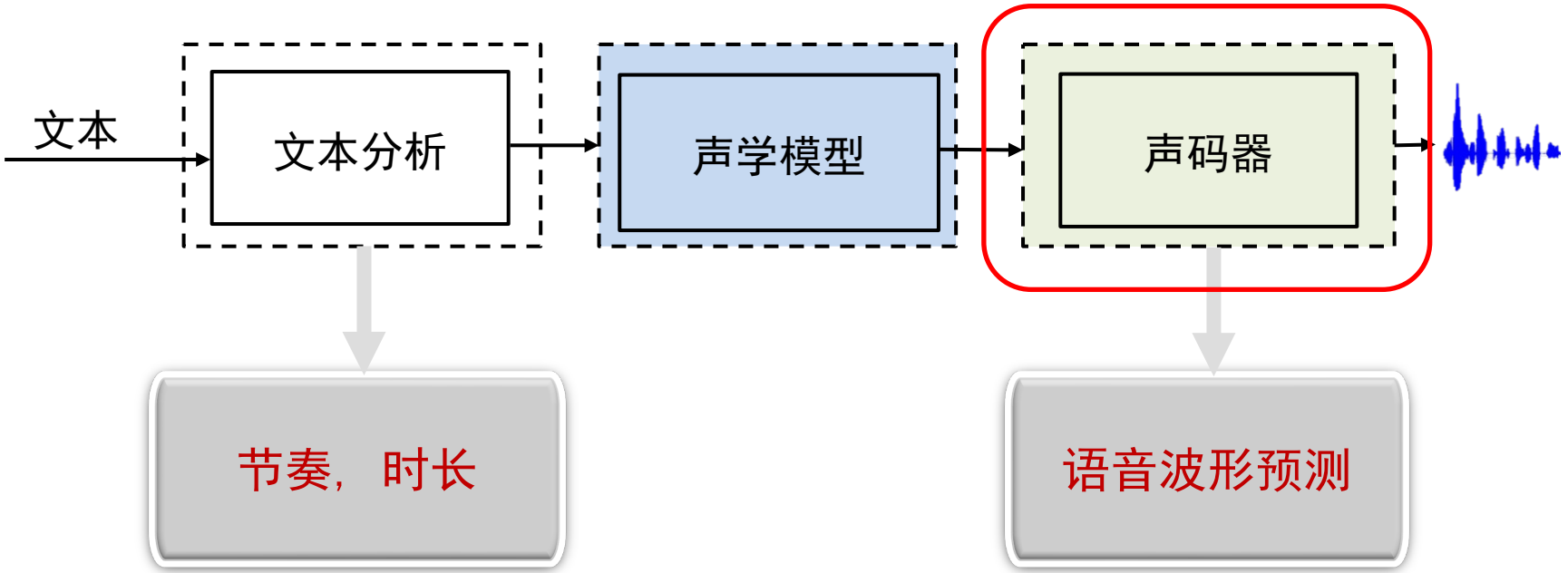
# 基于LSTM-RNN的声学模型

- 在不同声学模型下，对合成语音进行主观评价

| Model | # of params | 5-scale MOS |
|---|---|---|
| DNN | 3 749 79 | $3.370 \pm 0.114$ |
| LSTM-RNN | 476 435 | $3.723 \pm 0.105$ |

Zen, Heiga, and Haşim Sak. "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis." ***2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***. IEEE, 2015.

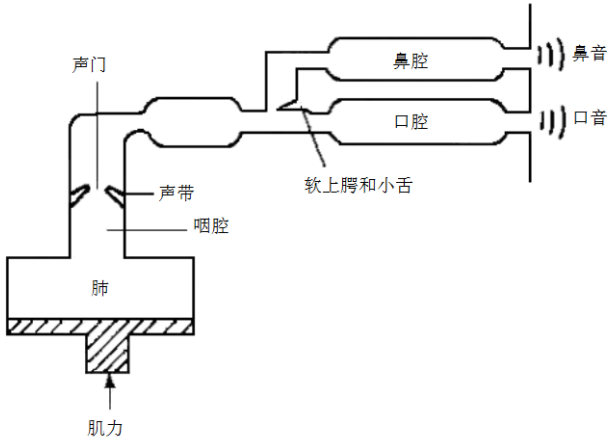
# 管道式语音合成



# 声码器

- 目的
  - 声码器实现声学参数到语音波形的转化。
- 基于信号处理的声码器
  - LSP/LSF
  - WORLD
  - STRAIGHT

# 传统语音声码器




基音周期

冲激序列发生器

声门脉冲模型 G(z)

Av

浊音/清音开关

随机噪声发生器

An

声道模型 V(z)

辐射模型 R(z)

产生的语音

# 共振峰声码器

## 共振峰合成器

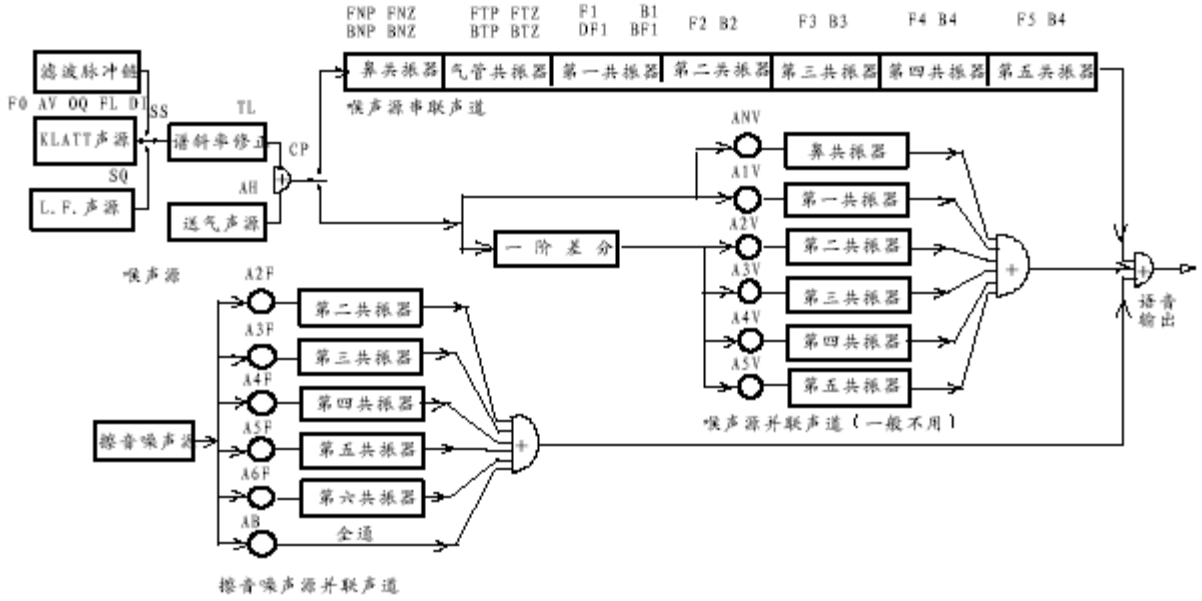
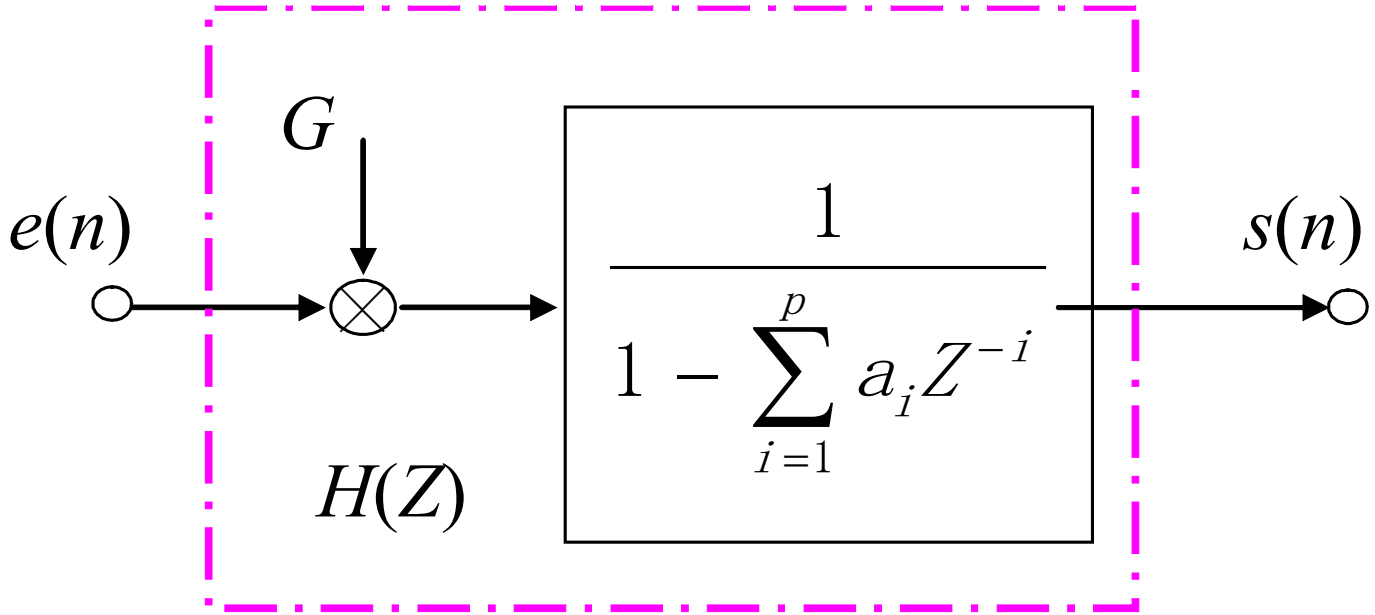



图 3. KLATT 共振峰语音合成器 （引自 Klatt D. 1987）

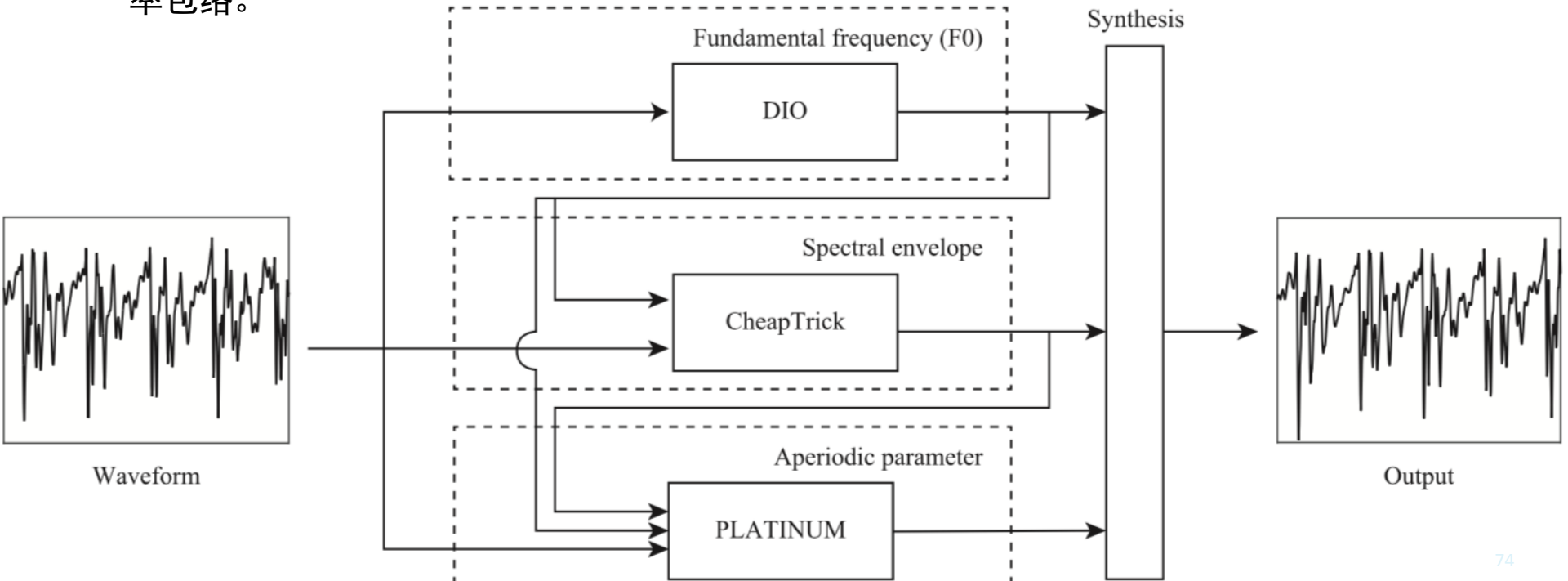
# 全极点模型



# WORLD声码器（目前是常用模型）

可以实现实时、高质量的语音合成。比传统的系统速度快10倍以上，RTF(real time factor)表明它可以应用于实时系统。通过3个语音信号相关的参数合成语音－基频f0， 谱包络spectrum envelope, 非周期信号参数aperiodic parameter。

1. 输入wave通过DIO算法估计出**F0 contour**（基频）
2. F0和wave作为输入，由cheap trick估计出**spectral envelope**(频谱包络)
3. 输入F0/sp/wave，用PLATINUM将提取出来的信号进行估计，得到**aperiodic parameter**(非周期参数)。非周期参数的定义和之前的不一样。
4. WORLD不能和STRAIGHT的衍生系列一样操纵非周期参数，但是可以和它们一样操作F0和频率包络。



# 管道式语音合成优缺点

- 优点
  - 语音合成过程清晰可解释，出了问题容易找到解决办法
  - 总体音质基本满足应用
- 缺点
  - 流程繁复
    - 包括文本分析，时长预测，声学模型和声码器模块；误差累计，优化目标不一致基于信号处理的声码器
  - 文本分析
    - 标注成本高昂，需要专家知识
  - 声学模型
    - 文本特征和声学特征通过HMM强制对齐
  - 声码器
    - 基于信号处理的声码器，不够自然

# 谢谢！