Chapter 13

Speech Synthesis
语音合成

主要内容

- 语音合成基本概念
- 语音合成系统构成
- 语音合成后端方法
- 统计参数语音合成方法
- 语音合成技术未来发展方向
- 合成样例演示

语音合成基本概念

基本概念

- 语音合成（Speech Synthesis）
  - 赋予机器像人一样自如说话的能力
- 典型的交叉学科
  - 涉及语言学、语音学、自然语言处理、信号处理、模式识别等
- 语音合成的三个层次
  - Text-to-Speech
  - Concept-to-Speech
  - Intention-to-Speech

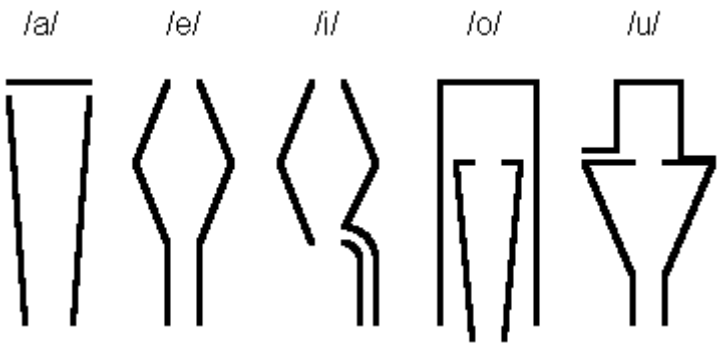
语音合成的应用

- 对话系统
  - Siri, 讯飞语点
- 电话信息查询
  - 话费查询、考试结果查询、股票交易查询等
- Eyes-Free应用
  - 车载导航、有声电子书
- 语言学习
  - 电子词典、单词例句朗读
- 实时信息播报
- 娱乐
- 视力/语言障碍者的信息获取与交流

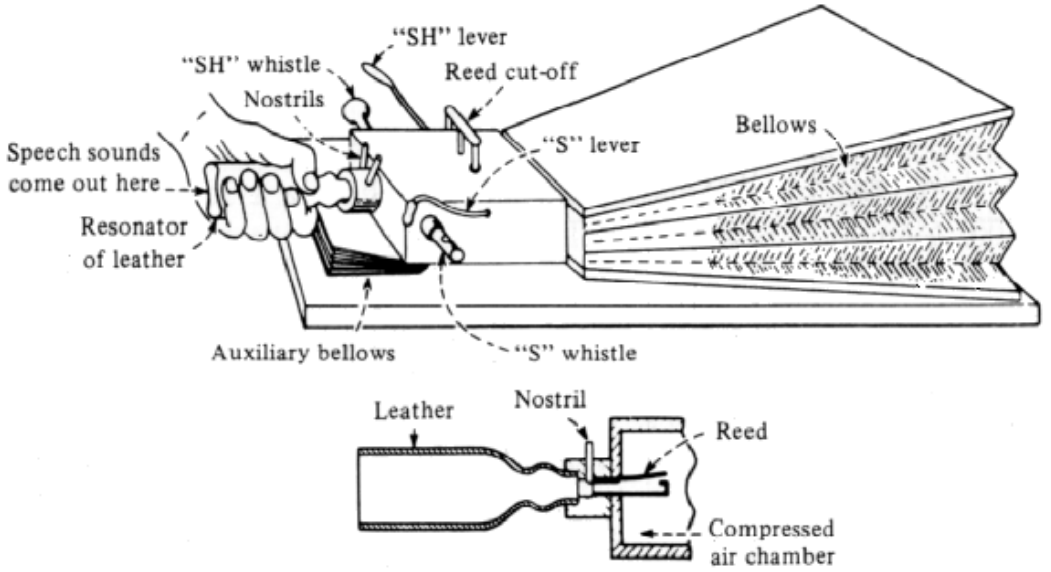
… …

语音合成的研究历史

- 机械装置合成
  - 18～19世纪



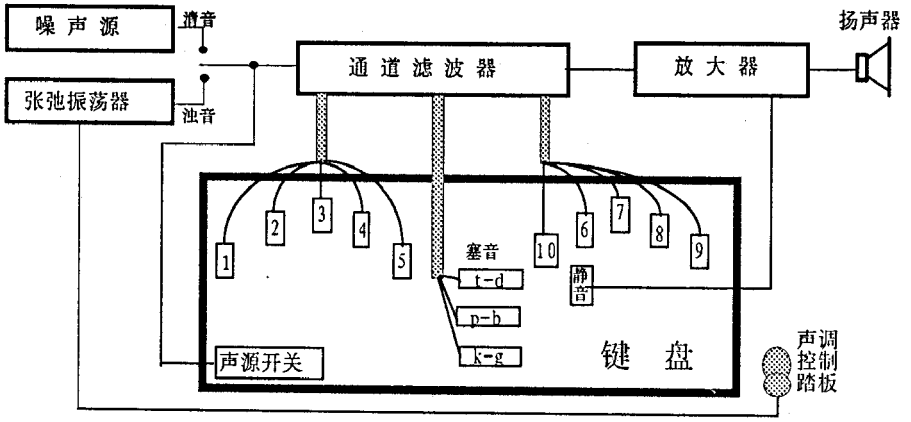
Kratzenstein’s Resonator(1779)



von Kempelen's speaking machine (1791)

语音合成的研究历史

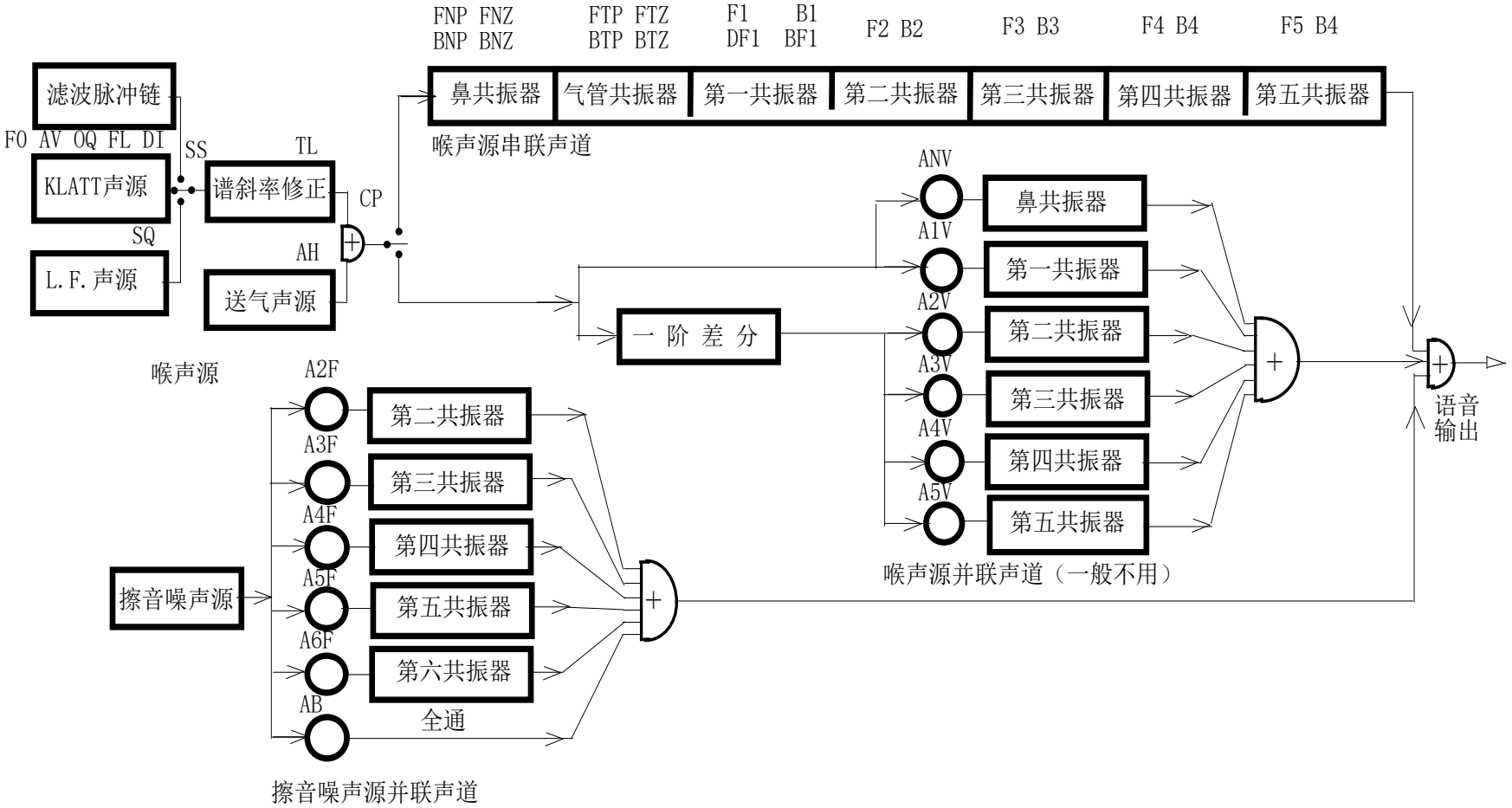
- 电子合成器
  - 1939，VODER，Bell Lab, Dudley



图一 "Voder" 简图(引自 Dudley H. 1939)

语音合成的研究历史

- 共振峰参数合成器＋规则合成
  - 1980，KLATT



语音合成的研究历史

- 波形拼接语音合成
  - 80年代末，PSOLA方法的提出
  - 90年代初，ATR提出大语料库语音合成方法

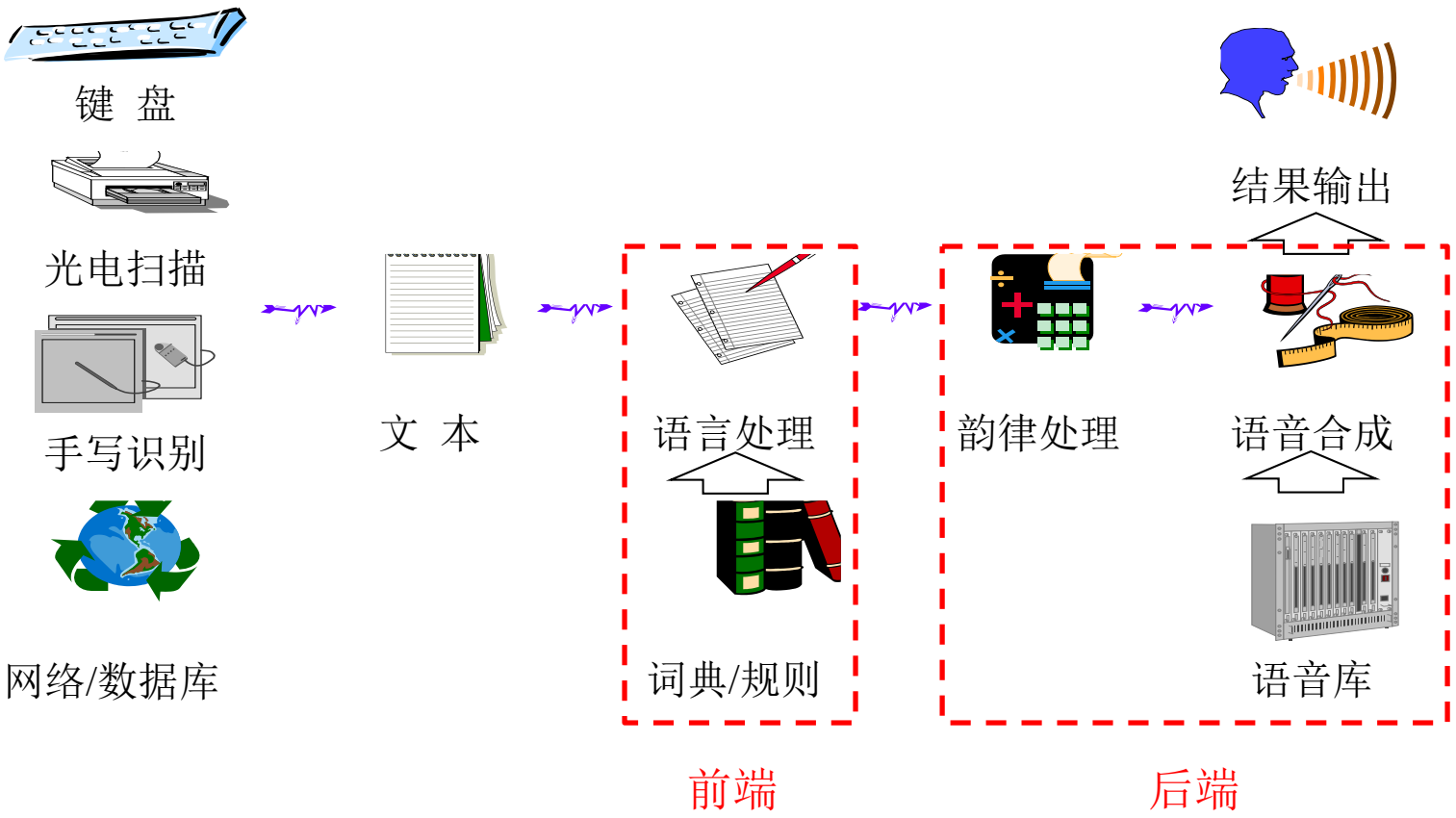
- 统计参数语音合成
  - 20世纪末，以基于HMM（隐马尔科夫模型）的参数合成方法为代表
  - 2013年后，基于神经网络的统计参数语音合成

语音合成性能评价

- 可懂度
  - 能否准确传达文本信息
  - 听写正确率(Diagnostic Rhyme Test)
- 自然度
  - 合成语音是否像真人一样自然流畅
  - Mean Opinion Score (1~5) / Preference Test
- 相似度
  - 合成语音音色是否接近目标人
  - Mean Opinion Score (1~5) / Preference Test

语音合成系统构成

语音合成系统构成



语音合成系统前端

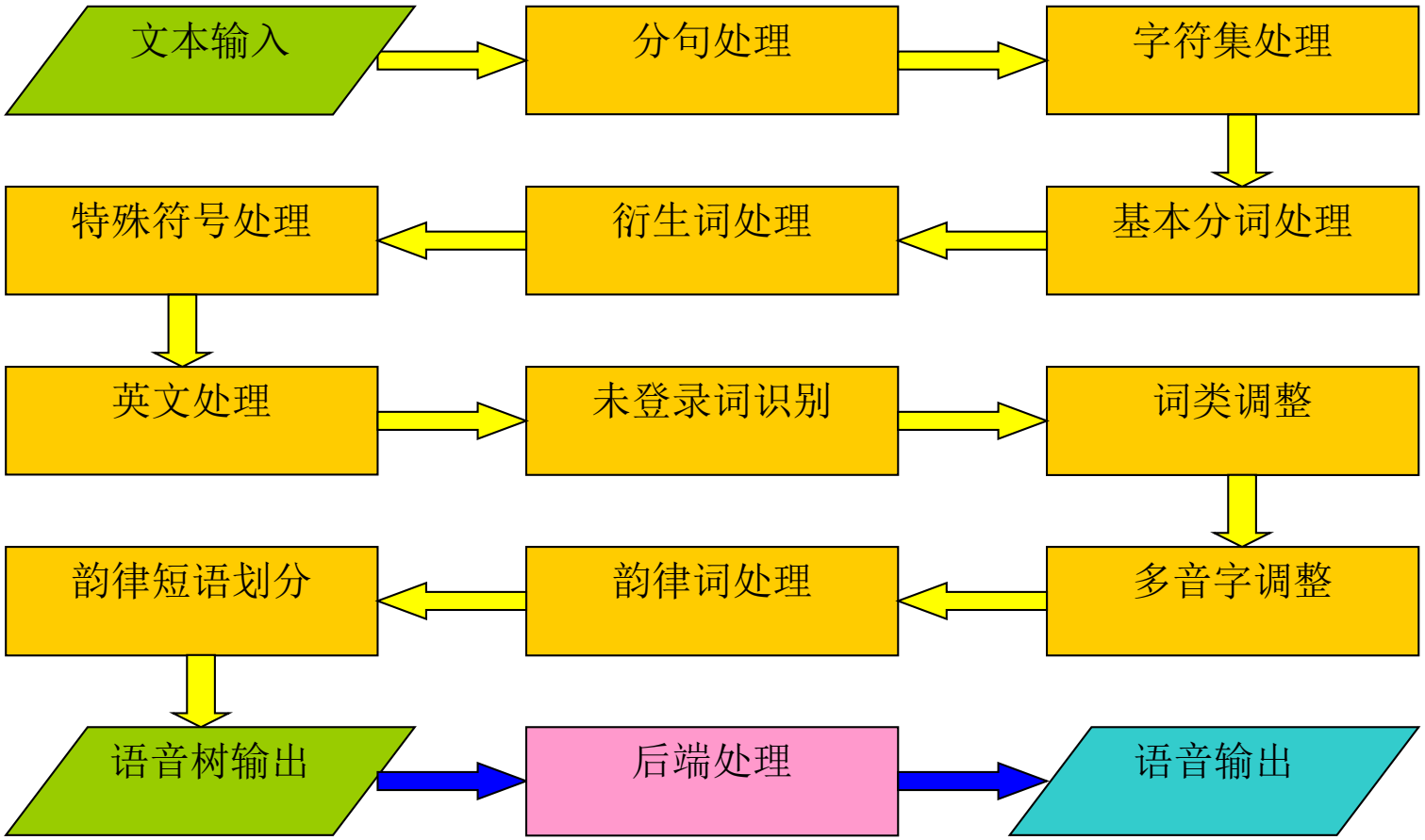
- 语音合成前端（Front-End）
  - Text Analysis
  - NLP(Natural Language Processing)研究范畴

- 目的
  - 对输入文本在语言层、语法层、语义层的分析
  - 将输入的文本转换成层次化的语音学表征
    - 包括读音、分词、短语边界、轻重读等
    - 上下文特征（context feature）

语音合成系统前端

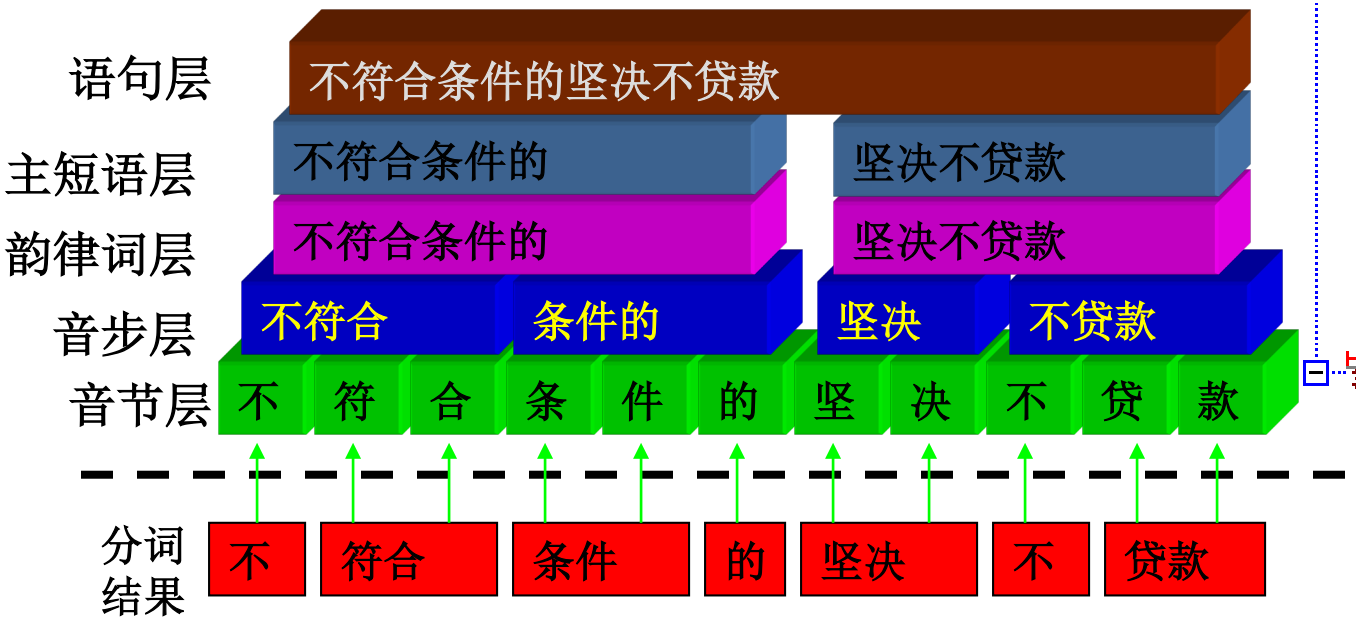
- 难点
  - 语种相关性
  - 语言现象之间的存在冲突，通过不断的资源修正与统计方法来解决或缓解
    - 多音字
    - 特殊符号
      - Mr, Dr, Rd
      - £5, $5 million, 12°C
      - 1995 2001 1,995 ☎236 3017 233 4488

中文语音合成系统前端典型处理步骤

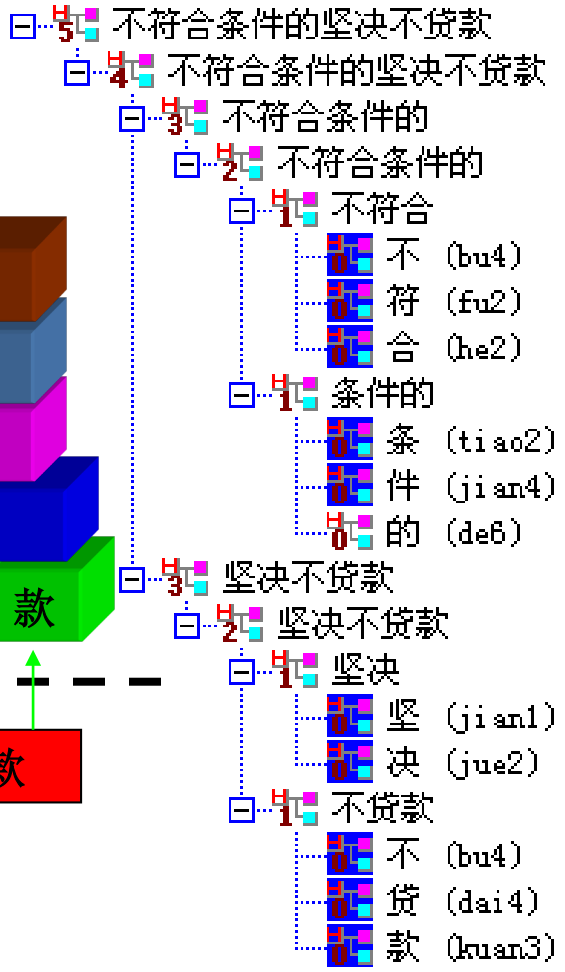


中文语音合成系统前端分析结果示例

例句：“不符合条件的坚决不贷款”



系统生成：



语音合成系统后端

- 语音合成后端（Back-End）
  - DSP(Digital Signal Processing)研究范畴
- 目的
  - 基于前端给出的层次化语音表征来生成语音
- 现阶段主要方法
  - 单元挑选与波形拼接
    Unit selection and waveform concatenation
  - 统计参数语音合成
    Statistical parametric speech synthesis

语音合成后端方法

单元挑选与波形拼接方法

- 技术实现
  - 从录制的语音数据库中选择合适的语音片段
  - 将各语音片段波形拼接得到最终合成语音
- 如：希望合成“中国科大”，可能取以下录音片段
  - 他说话声如洪钟
  - 较强能空气将影响我国中东部
  - 科学技术是第一生产力
  - 大家都不要讨论别人的是非
- 发展历程
  - Single inventory: diphone synthesis [Moullnes; ’90]
  - Multiple inventory; unit selection synthesis (USS)
    - ATR v-Talk [Sagisaka; ’92], CHATR [Black; ’96]
    - AT&T Next-Gen TTS [Beutnagel; ’99]

单元挑选与波形拼接方法

- 关键问题
  1. 用什么类型的语音段作为挑选单元
  2. 单元库的设计与录制
  3. 如何进行合理的单元挑选
  4. 波形拼接时的信号处理

单元挑选与波形拼接方法

1. 单元尺度选择
   - 单元尺度长=>合成质量更高，对于单元库的规模要求更大

Table 16.4 Unit types in English assuming a phone set of 42 phonemes. Longer units produce higher quality at the expense of more storage. The number of units is generally below the absolute maximum in theory: i.e., out of the $42^3 = 74{,}088$ possible triphones, only about 30,000 occur in practice.

Unit length	Unit type	#Units	Quality
Short	Phoneme	42	Low
↓	Diphone	~1500	↓
	Triphone	~30K	
	Demisyllable	~2000	
	Syllable	~11K	
	Word	100K–1.5M	
	Phrase	∞	
Long	Sentence	∞	High

单元挑选与波形拼接方法

2. 单元库的设计与录制
   - 考虑每个单元及其在不同上下文环境中出现的频度进行录音语句的挑选
   - 通常由专业播音人员进行中性风格的语音录制 (>1hr)
   - 录制数据库的标注
     - 音段标注（音素序列、音素边界）
     - 韵律标注（韵律层级）
     - 人工或自动进行

单元挑选与波形拼接方法

3. 单元挑选方法

- 寻找最优的备选单元序列，使其与目标单元序列尽量匹配
- 基于代价函数进行

$$d(\Theta,T)=\sum_{j=1}^{N} d_u(\theta_j,T)+\sum_{j=1}^{N-1} d_t(\theta_j,\theta_{j+1})$$

$$\hat{\Theta}=\arg\min_{\Theta} d(\Theta,T)$$

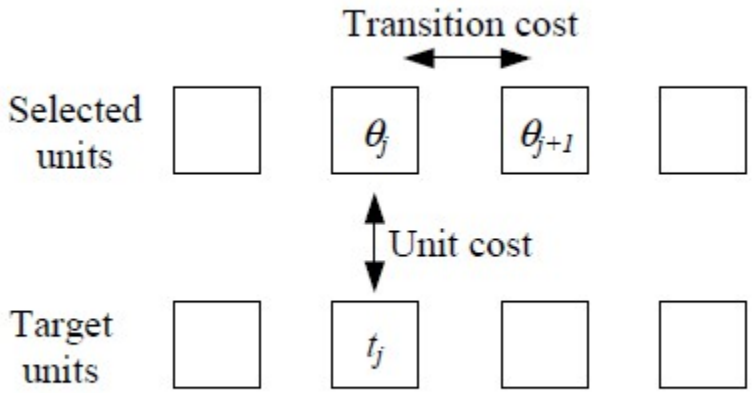


Figure 16.6 Tradeoff between unit and transition costs.

- 目标代价(unit /target cost)：衡量备选单元与目标单元匹配程度
- 连接代价(Transition/joint/concatenation cost)：衡量前后两个备选单元连接时的平滑程度

单元挑选与波形拼接方法

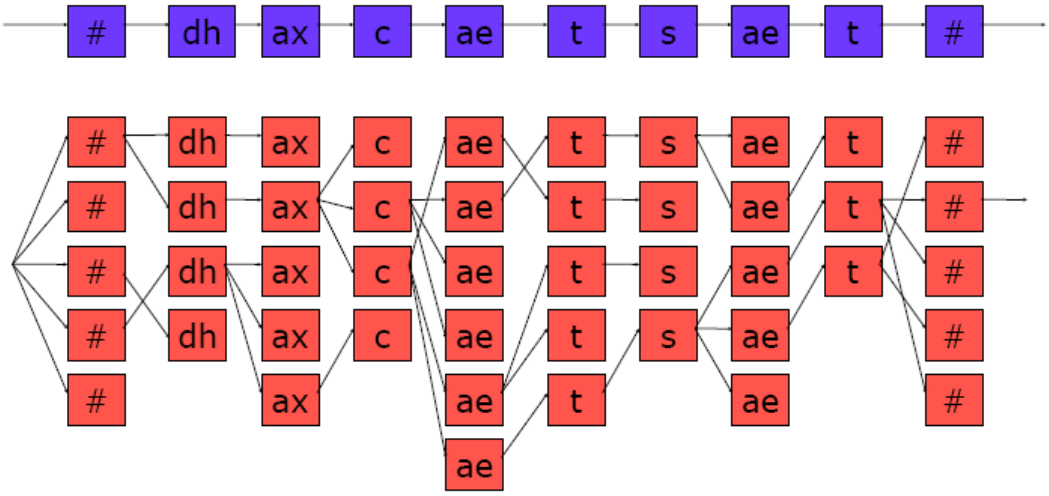
3. 单元挑选方法

– 代价函数设计
  - 基于经验设计代价表
  - 备选单元与预测目标单元声学特征的距离
  - 基于统计声学模型的概率准则

Table 16.5 Cost matrix for intrasyllable concatenations (after Yi [55]). The rows represent the left side of the transition and the columns represent the right side, and NA represents a case that does not occur.

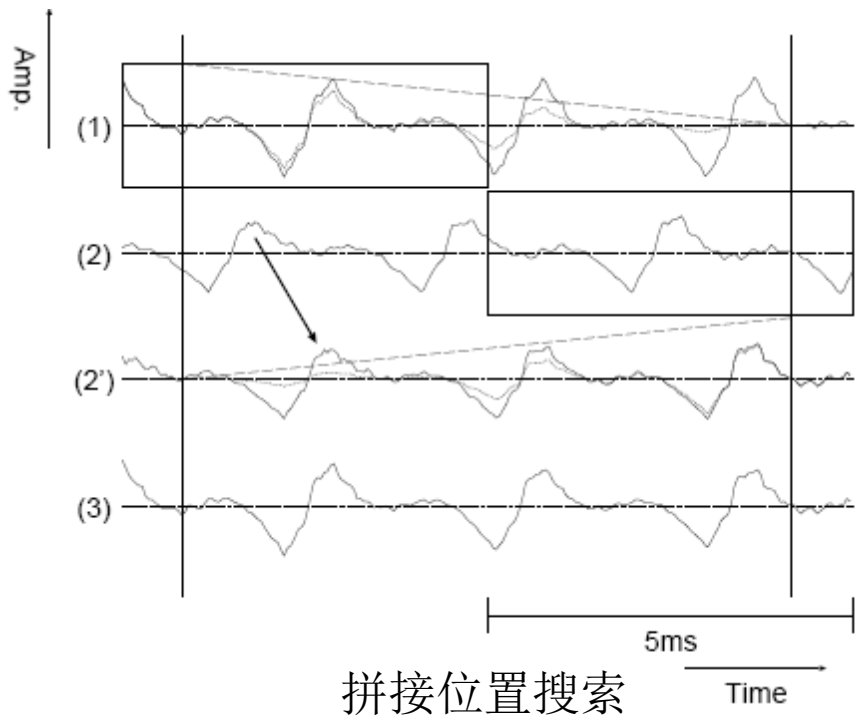
	vowel	semivowel	nasal	obstruent	/h/
vowel	10,000	10,000	7500	10	NA
semivowel	10,000	7500	7500	10	NA
nasal	5000	10	NA	10	NA
/h/	5000	NA	NA	NA	NA
obstruent	10	10	10	10,000	NA

– 最优序列搜索算法
  - 利用Viterbi Search降低搜索运算量
  - 保存每一步中，以每一个备选单元结尾的部分最优路径



单元挑选与波形拼接方法

4. 波形拼接时的信号处理
   - 备选单元的能量调整
   - 基于波形相关度的最佳拼接位置搜索
   - 前后备选单元波形的加窗叠加
   - 修改语音段的韵律特征（时长、基频）使之与目标的韵律特征相匹配（可能）——PSOLA



统计参数语音合成方法

- 技术实现
  - 训练阶段
    - 利用声码器提取训练语音数据库中语音波形的声学特征参数
    - 训练统计声学模型$P(O|C)$，其中O为声学特征，C为上下文特征
  - 合成阶段
    - 由前端文本分析得到带合成语句对应的上下文特征
    - 基于统计声学模型预测该上下文特征对应的最优声学特征
    - 将预测的声学特征送入声码器重构语音信号
- 发展历程
  - Proposed in mid-'90s, becomes popular since mid-'00s
  - Large data + automatic training
    ⇒ Automatic voice building
  - Source-filter model + statistical acoustic model
    ⇒ Flexible to change its voice characteristics
  - HMM as its statistical acoustic model
    ⇒ HMM-based speech synthesis (HTS) [Yoshimura; '99]

两种方法的对比

	单元挑选与波形拼接	统计参数合成
合成自然度	较高但不平稳	平稳，不错
合成音质	很好	有点低
表现力	受音库录音风格限制	相对中庸些，但具有较强变化能力
与原始发音人相似度	很相似	因过平滑而不完全相似
系统大小	很大，2G	很小，1M
系统研发难度	很难	很快捷
多语种能力	很难，完全各自开发	后端语种无关性好
合成效率	效率较高，但需要很大内存	因为涉及较多复杂模型运算，CPU消耗较大
拓展能力	无	可衍生唱歌合成、个性化合成等

统计参数语音合成方法

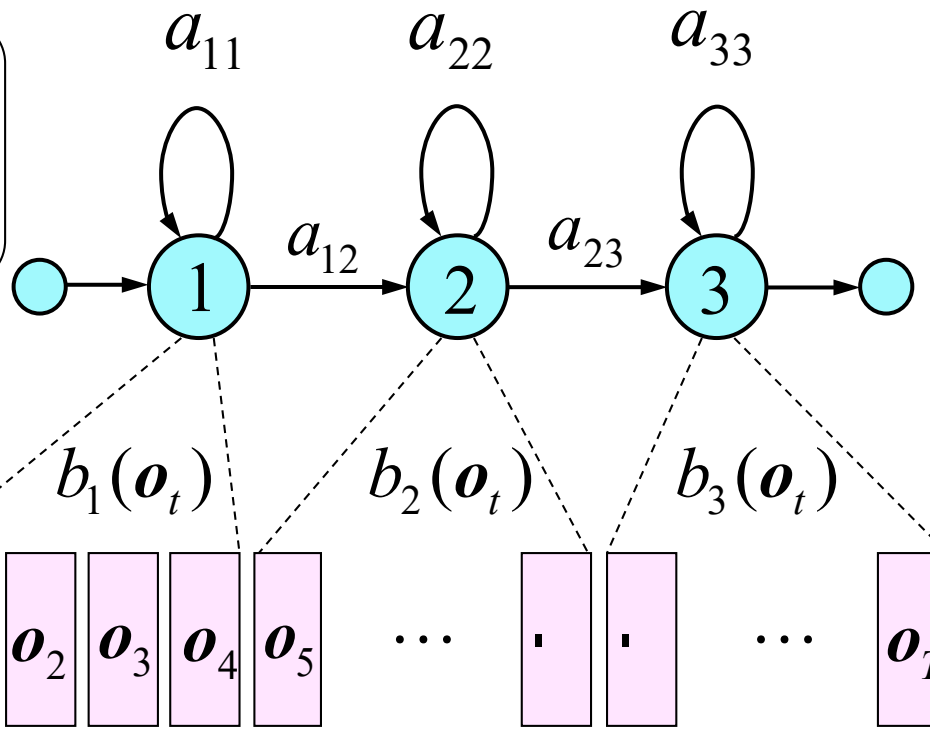
Hidden Markov Model

- 语音是由一组特定的发音方式形成的声音编码波形。我们（计算机）所能观测到的是语音的波形及其所提取的特征。这些特征与具有特定意义的发音以某种关系联系起来的
- 在HMM语音模型中，假设一组具有特定意义的发音为一个Markov链，每一个不同发音方式看成此马尔可夫链的状态。
  - 一个音素作为HMM模型：则状态定义为发音的方式或过程
  - 一个汉字作为HMM模型：则每个状态可定义为音素、半音节等
- 由于Markov的状态不能被直接观测到，所能观测到的是与状态以某种概率关系联系起来的观测量（即语音信号或语音特征），故称这种Markov模型是隐含的，这就是隐含马尔可夫模型。

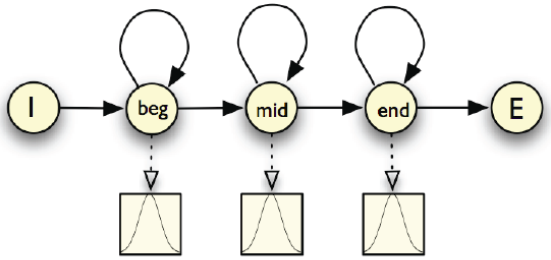
Hidden Markov Model

a_{ij} : 状态转移概率

$b_q(\boldsymbol{o}_t)$: 输出概率

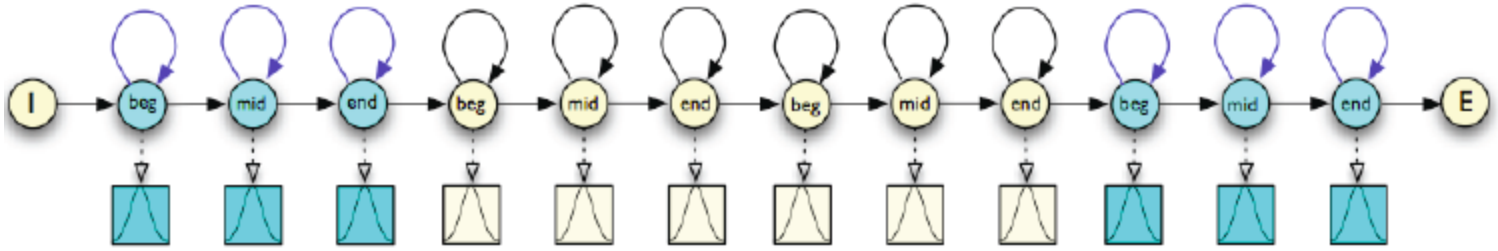


HMM声学模型



音素HMM

/ih/



/s/ /ih/ /k/ /s/

"six"

单词/句子HMM

基于HMM的参数语音合成

- 建模单元
  - 考虑最简单情况——单音素(monophone)
- 模型训练
  - 需要估计每个单音素HMM中的a_{ij}，以及各状态$b_q(\boldsymbol{o}_t)$中的分布参数
  - 从音库中的第*n*句训练语句中提取
    - 声学特征序列 $\mathbf{o}_n$
    - 音素序列=>句子级HMM λ_n
  - 最大似然(Maximum Likelihood)：通过最大化$\prod_{n=1}^{N} P(\mathbf{o}_n \mid \lambda_n)$实现对于所有单音素HMM对应模型参数的估计

基于HMM的参数语音合成

- 参数生成
  - 利用前端文本分析结果，得到待合成句的音素序列
  - 基于已训练的单音素HMM，拼合得到待合成句HMMλ
  - 最大化模型的输出概率实现参数生成

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

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

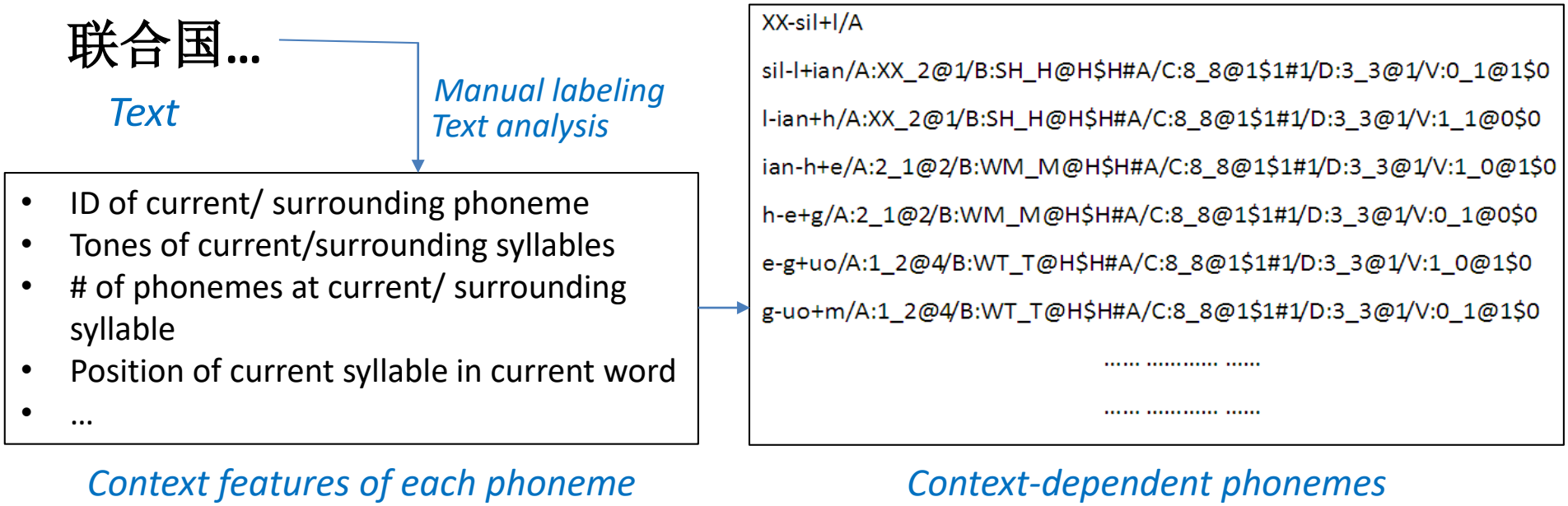
$$\hat{\boldsymbol{q}}=\arg\max_{\boldsymbol{q}}P(\boldsymbol{q}\mid w,\boldsymbol{\lambda})$$ 状态时长模型决定

$$\hat{\boldsymbol{o}}=\arg\max_{\boldsymbol{o}}P(\boldsymbol{o}\mid\hat{\boldsymbol{q}},\boldsymbol{\lambda})$$ 输出概率决定

基于HMM的参数语音合成

- 具体实现
  - 声码器与声学特征选择
    - 声码器：STRAIGHT(去除基频周期性对于频谱特征提取影响)
    - 激励源：基频、周期噪声比
    - 滤波器：LSP, mel-cepstrum …
    - 加入动态声学特征：解决生成特征的不连续问题
  - 建模单元
    - monophone -> triphone -> context-dependent phone
  - 模型结构
    - context-dependent phone数目过多：基于决策树的模型聚类
    - 基频特征的清浊特性：多空间概率分布(Multi-Space Probability Distribution)
  - 训练准则
    - Maximum Likelihood (ML)
    - Minimum Generation Error (MGE)

基于HMM的参数语音合成



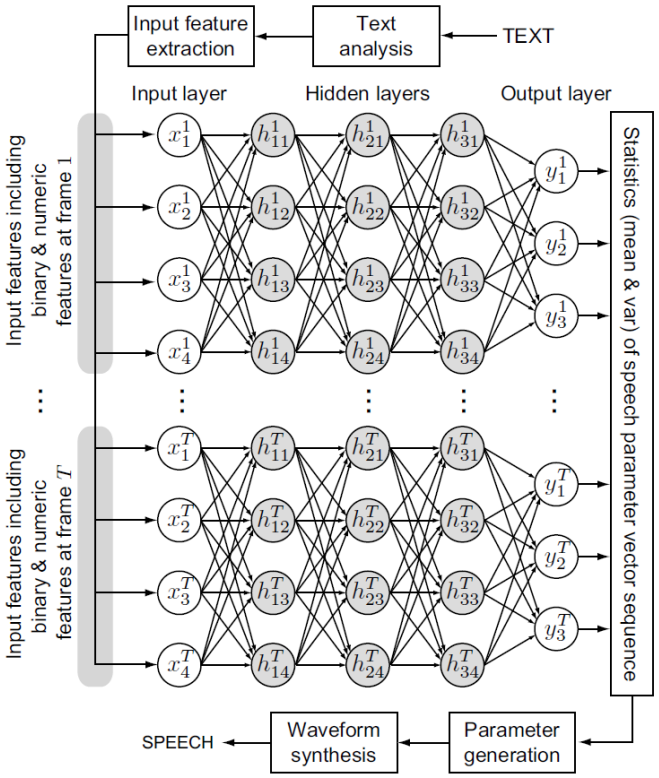
Softwares

HTS: Toolkit for HMM-based speech synthesis

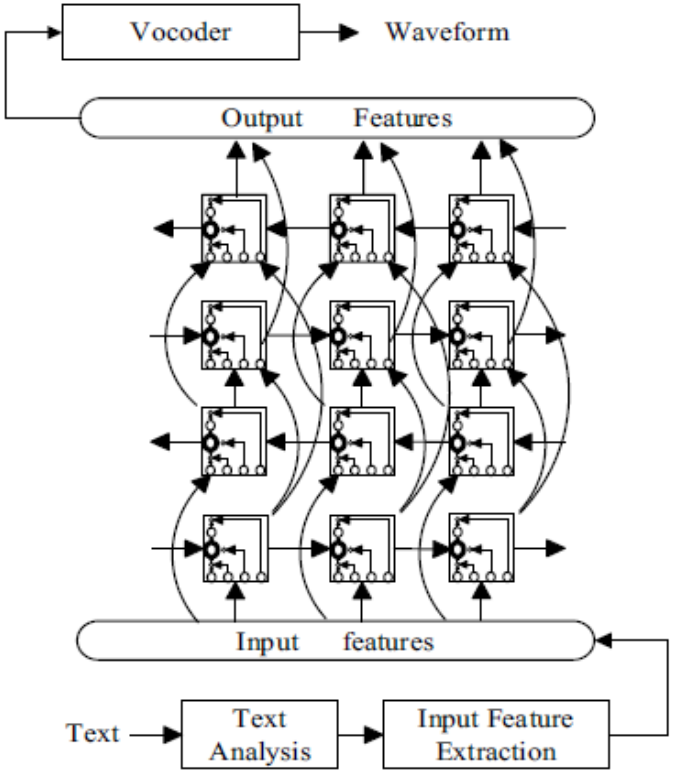
- Web: http://hts.sp.nitech.ac.jp/
- Research platform for HMM-based speech synthesis
- Released as a patch code for HTK
- Speaker dependent (SD) / adaptation (SA) demo scripts

基于神经网络的声学建模

- 基于深度神经网络(Deep Neural Network, DNN)和递归神经网络(Recurrent Neural Network, RNN)



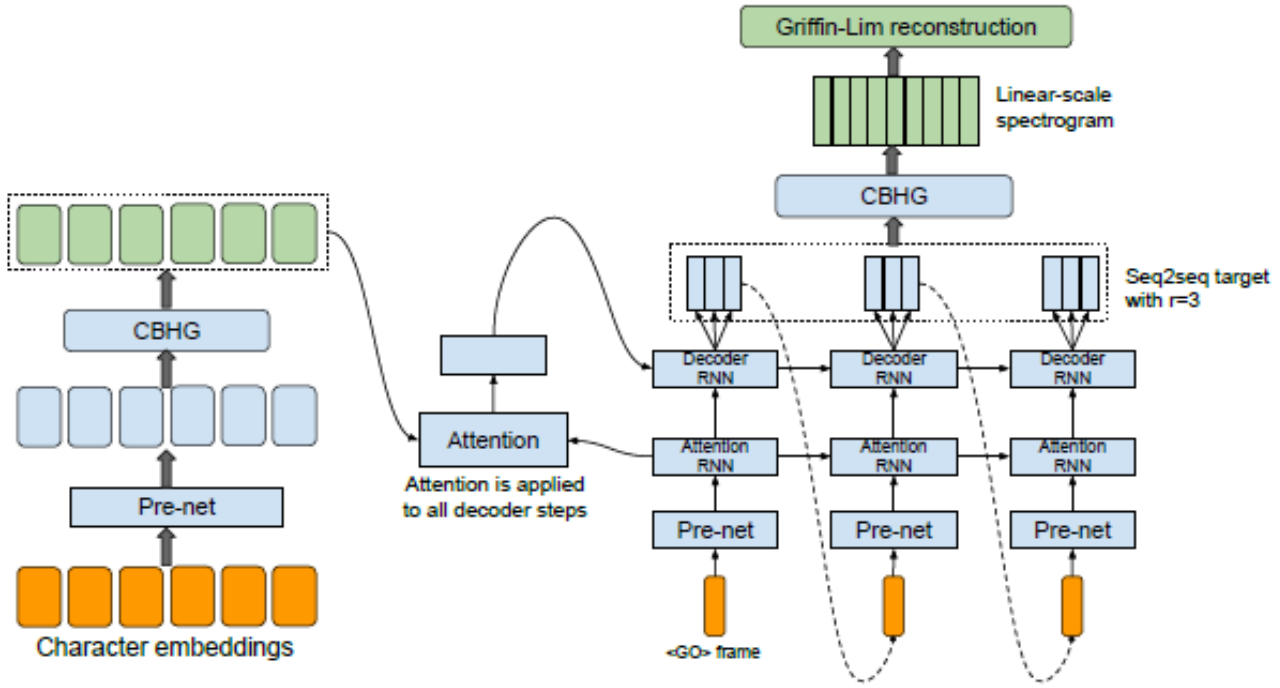
[Zen 2013]



[Fan 2014]

基于神经网络的声学建模

- 基于序列到序列(Sequence-to-Sequence)神经网络



Tacotron模型结构 [Wang 2017]

语音合成技术未来发展方向

发展方向

- 进一步提高合成语音自然度
  - 统计参数语音合成：音质受损与语调平淡问题
- 高灵活度的语音合成
  - 快速自适应：话者、口音、风格…
- 高表现力的语音合成
  - 丰富的语气语调与情感表现
- 多模态语音合成
  - 可视语音合成 Visual TTS
- TTS => CTS

Demo

单元挑选与波形拼接合成

科大讯飞语音合成系统

年份	1995年	1998年	1999年	2001年	2003年
自然度	<3.0	3.0	3.5	3.8	4.3





- 2009年
  - 中文合成系统
  - 英文合成系统

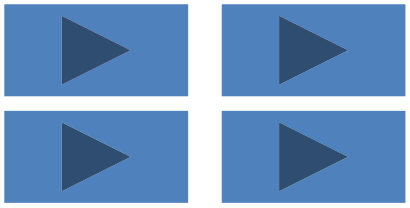




统计参数语音合成

- 统计参数语音合成
  - 基于HMM对语音进行建模，并通过训练得到合成所需的参数预测模型
  - 基本不需要人工干预的情况下自动、快速地进行系统构建
  - 合成语音具有很高的自然度
  - 音质相比拼接合成尚有一定差距

- 中文合成系统
- 英文合成系统



STOP

统计参数语音合成

- 话者自适应

源发音人	蜡笔小新	林志玲
	马三立	小丸子