

语音合成

洪青阳 副教授

厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

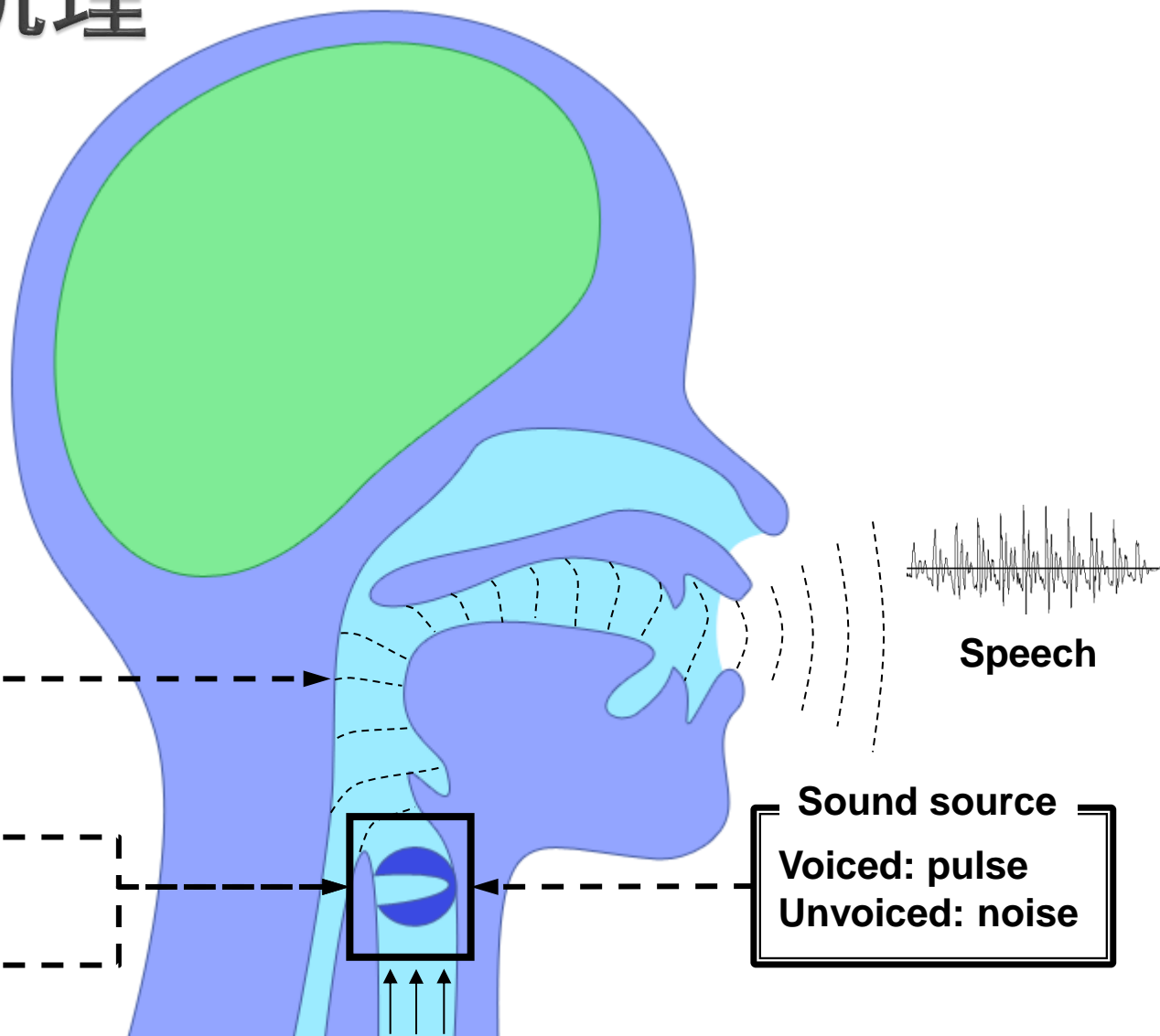
要点

- ▶ 语音产生机理
- ▶ 语音合成系统构成
 - 前端(Front-End)
 - 后端(Back-End)
- ▶ 语音合成后端方法
 - 参数合成法
 - 波形拼接法
 - 基于HMM的统计参数合成法
 - DNN合成法
- ▶ 语音合成发展方向

语音产生机理

Modulation of carrier wave
by speech information

- Frequency transfer characteristics
- Magnitude start--end
- Fundamental frequency

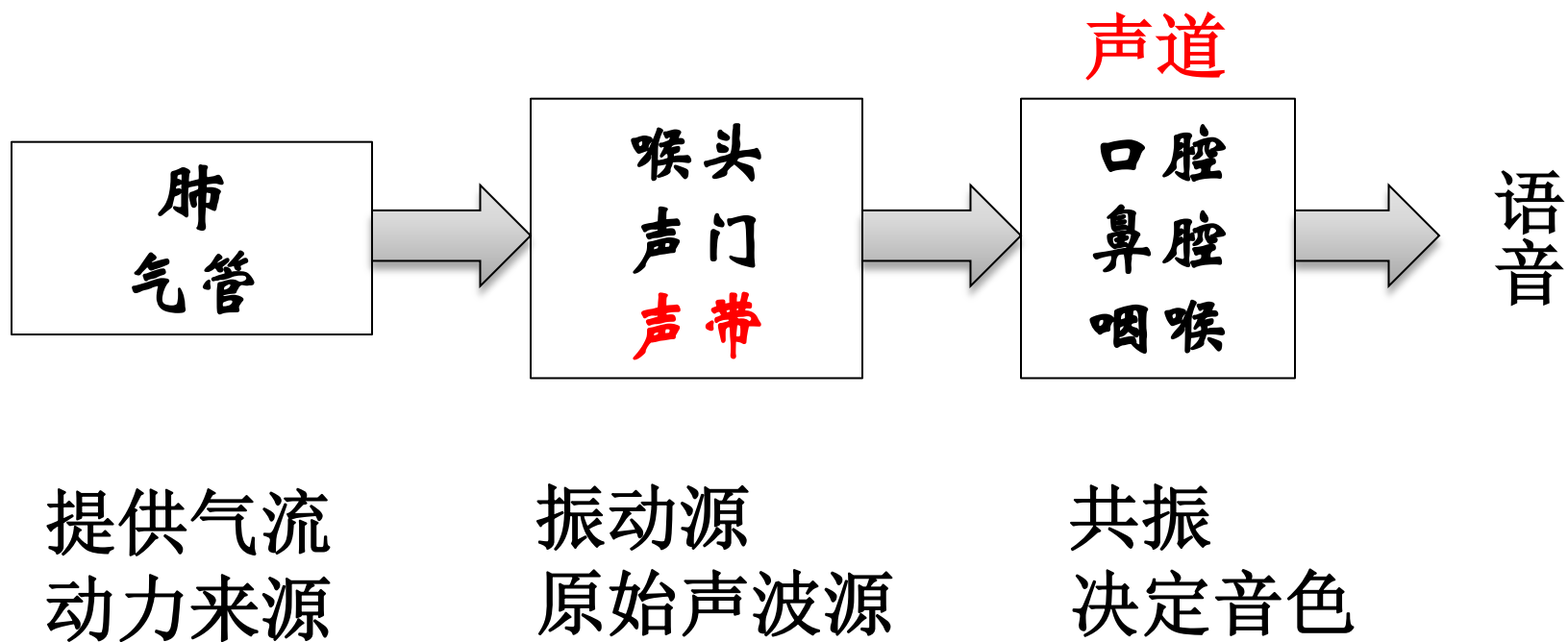


Sound source
Voiced: pulse
Unvoiced: noise

air flow

From HTS Slides released by HTS Working Group

语音产生机理



声道的作用

- ▶ 声道：截面非均匀的声管
- ▶ 形成复杂的共鸣体
- ▶ 改变声道形状改变共振峰，改变音色
- ▶ 共振峰（位置、带宽、幅度）决定元音音色

语音的物理特性

▶ 音色（共振峰）

[a] [i]

▶ 音高（频率）

▶ 音强（振幅）

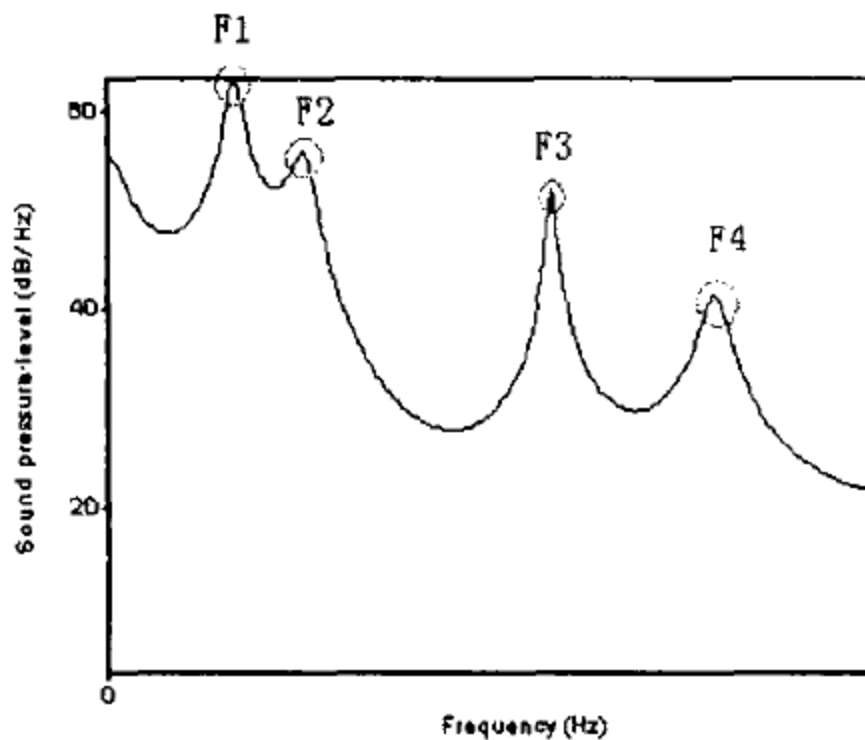
韵律特征

重音

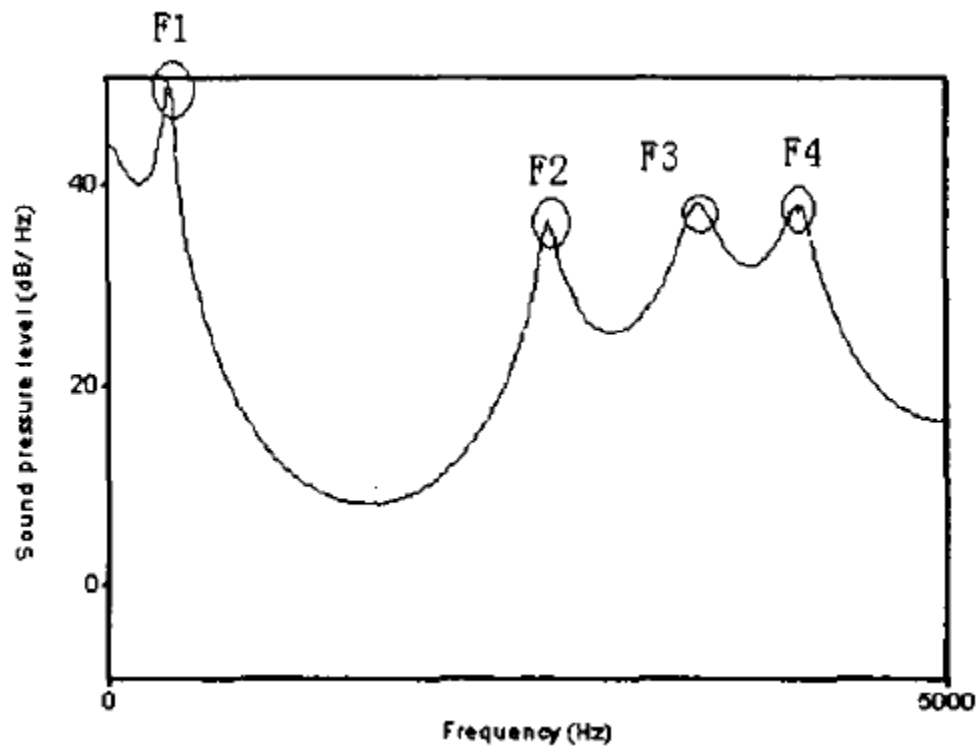
▶ 音长（延续时间）

[i] [i:]

共振峰



[a]



[i]

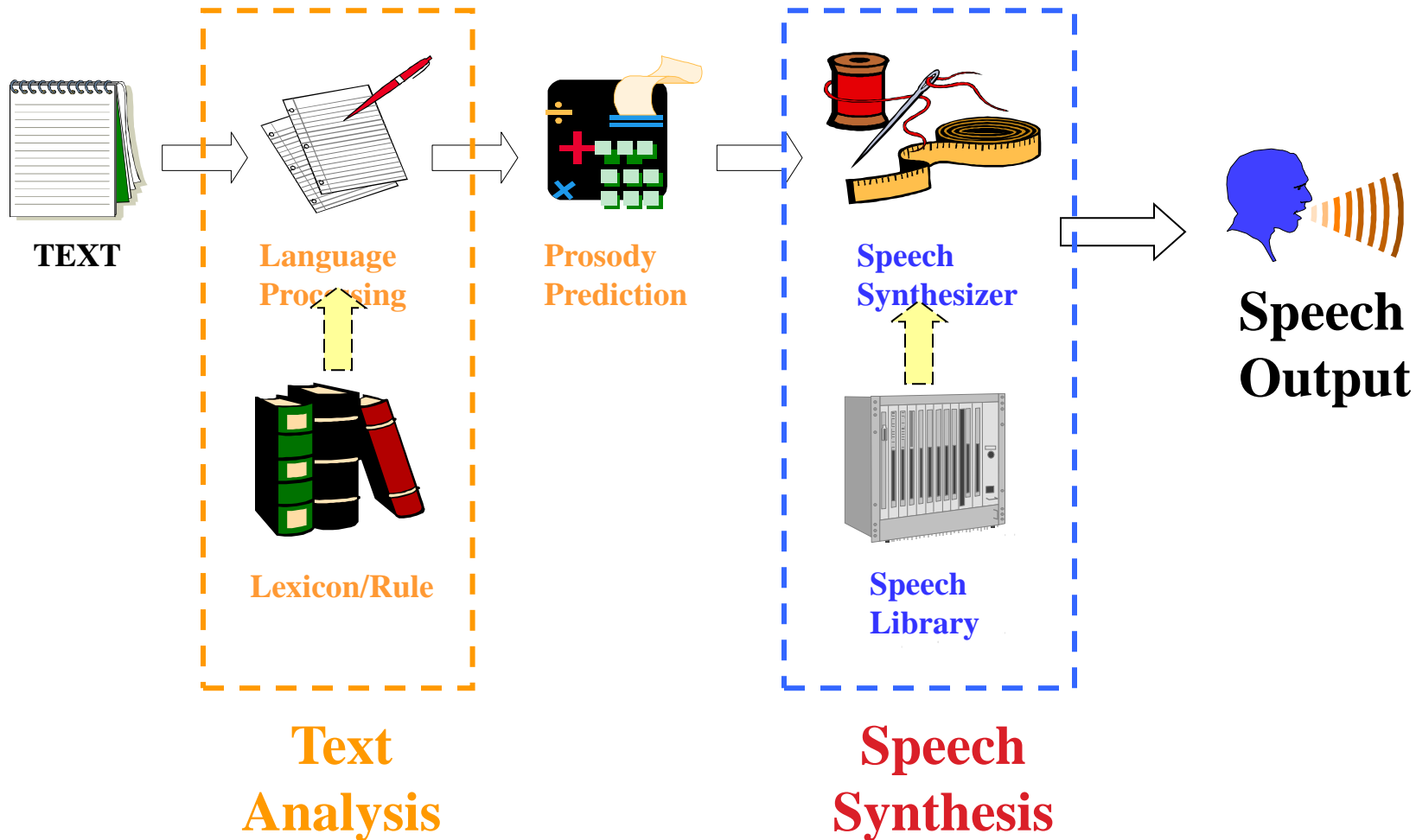
元音音色由共振峰的频率值和相对关系决定

音色与共振峰

- ▶ **低频共振峰强**
 - 音色混厚
- ▶ **中频共振峰强**
 - 音色圆润、自然
- ▶ **高频共振峰强**
 - 音色明亮、清透

语音合成系统构成

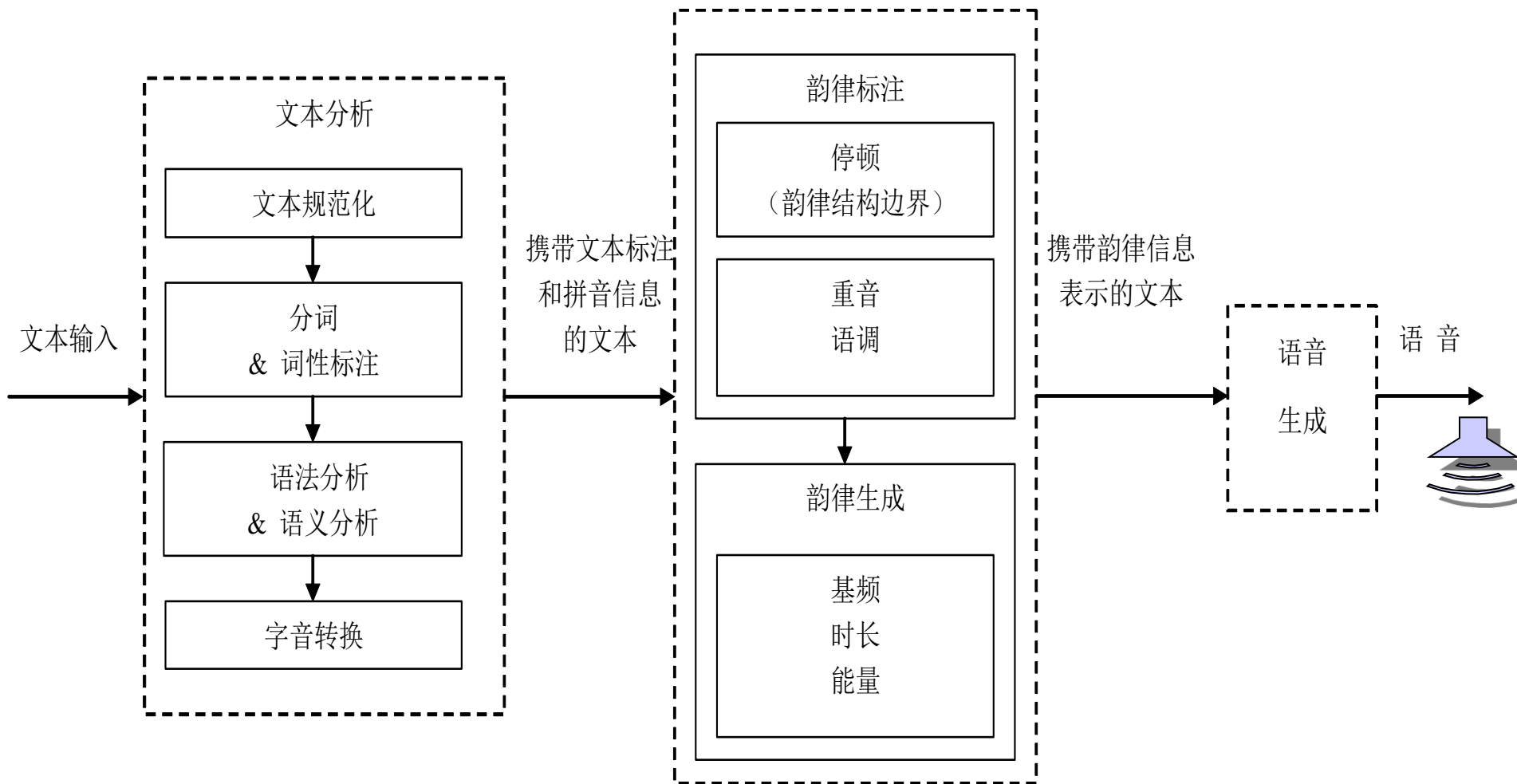
Text-to-Speech (TTS)



语音合成系统前端

- ▶ 语音合成前端(Front-End)
 - Text Analysis
- ▶ 目的
 - 对输入文本在语言层、语法层、语义层的分析
 - 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，以及上下文特征
- ▶ 难点
 - 多音字
 - 特殊符号，如3:20 PM
 - 语种相关性

语音合成系统前端处理



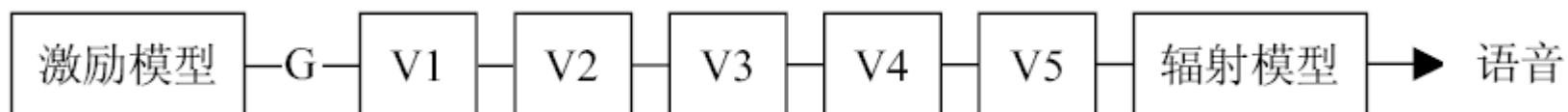
语音合成后端方法

- ▶ 参数合成法(~'90s)
 - 共振峰合成
- ▶ 波形拼接法('90s~)
- ▶ 基于HMM的统计参数合成法
- ▶ DNN合成法

共振峰合成法

- ▶ 级联型模型
- ▶ 并联型模型
- ▶ 混合型模型

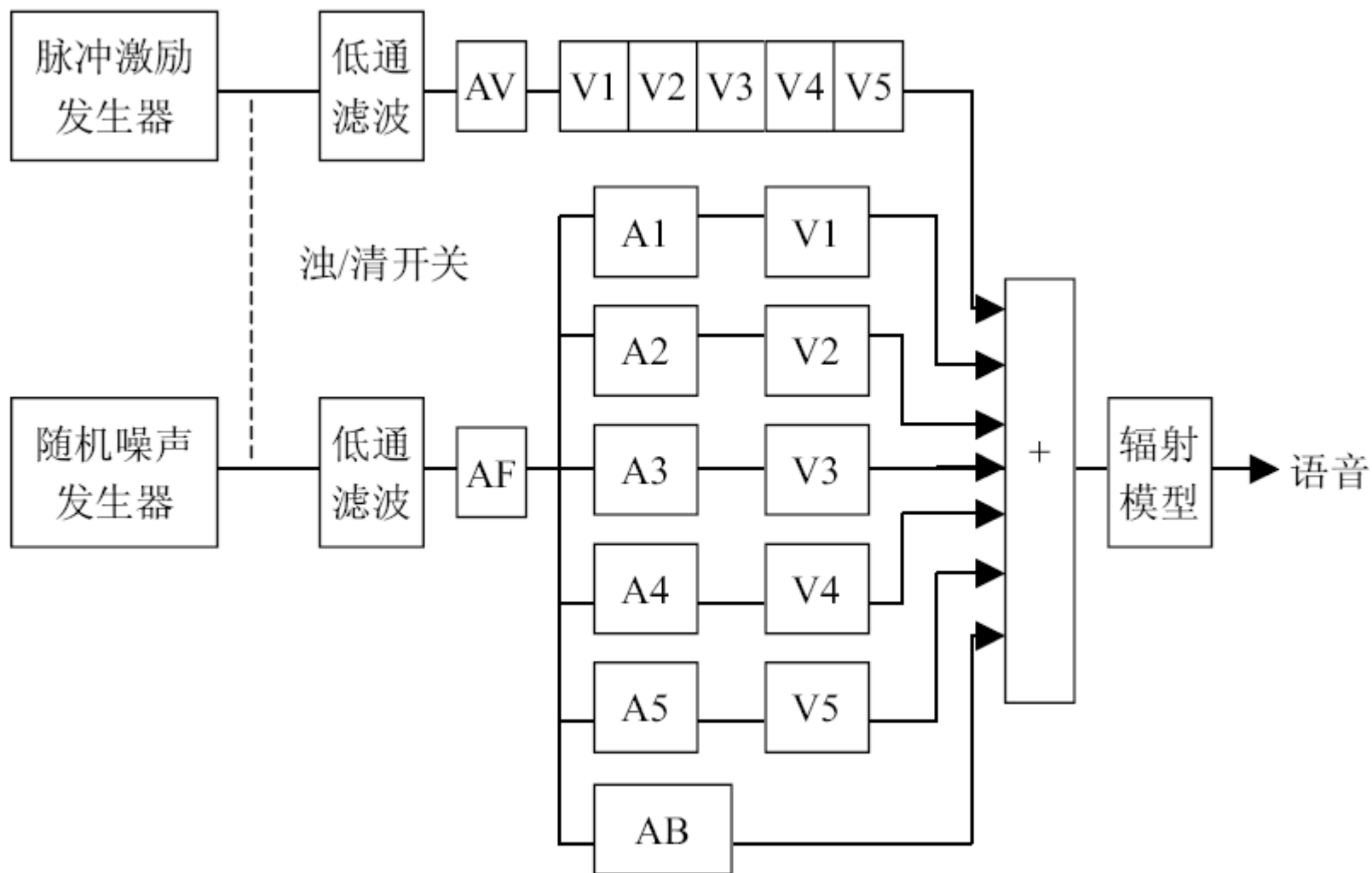
级联型共振峰模型



V1 V2 ...V5 对应于不同共振峰的滤波器

主要用于元音合成

混合型共振峰模型



共振峰合成方法的优缺点

▶ 优点

- 参数具有物理意义
- 语音宏观特性表达

▶ 缺点

- 模型粗糙，逼真度低
- 控制参数多，调整困难

波形拼接法

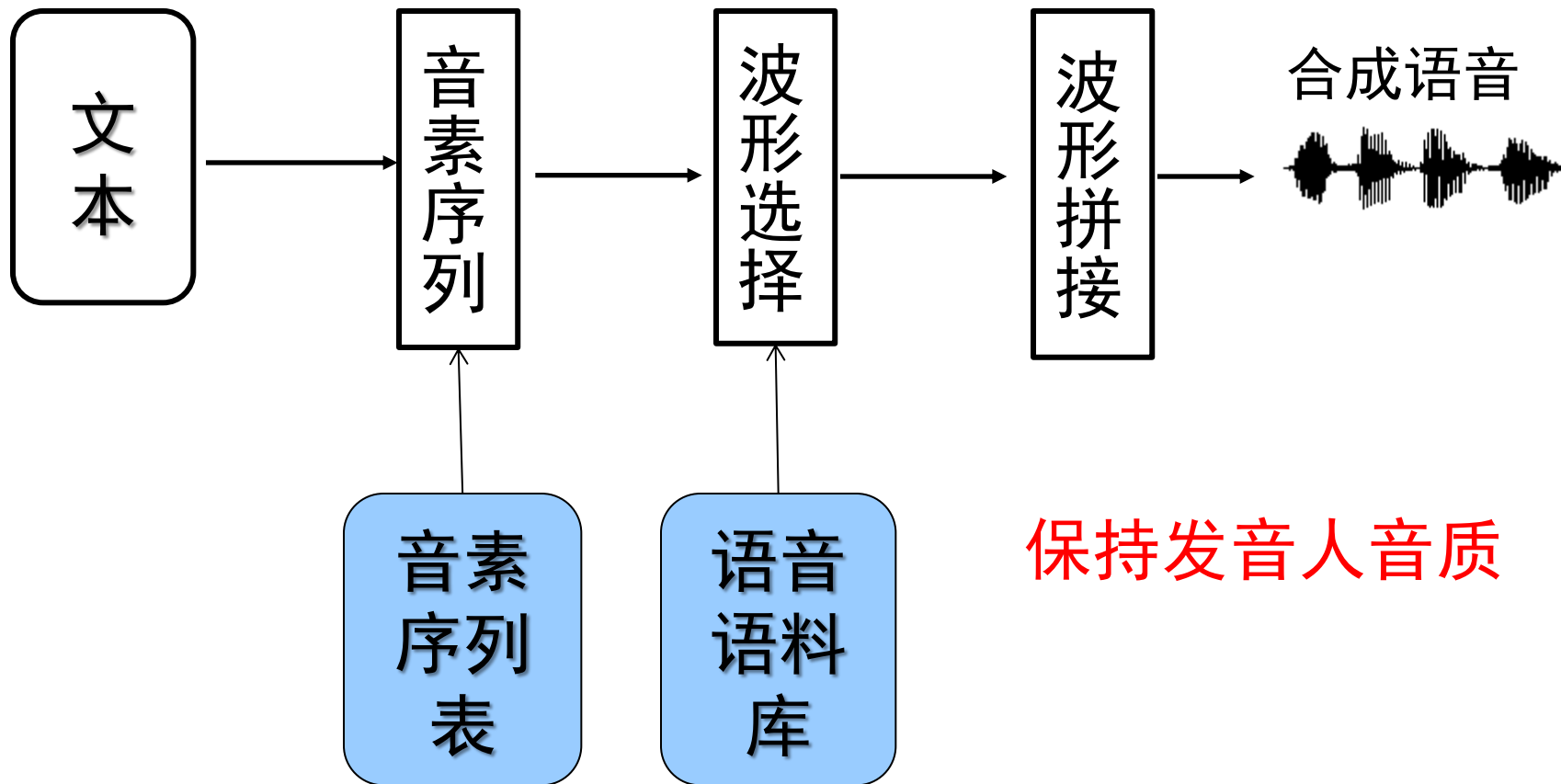
▶ 技术实现

- 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

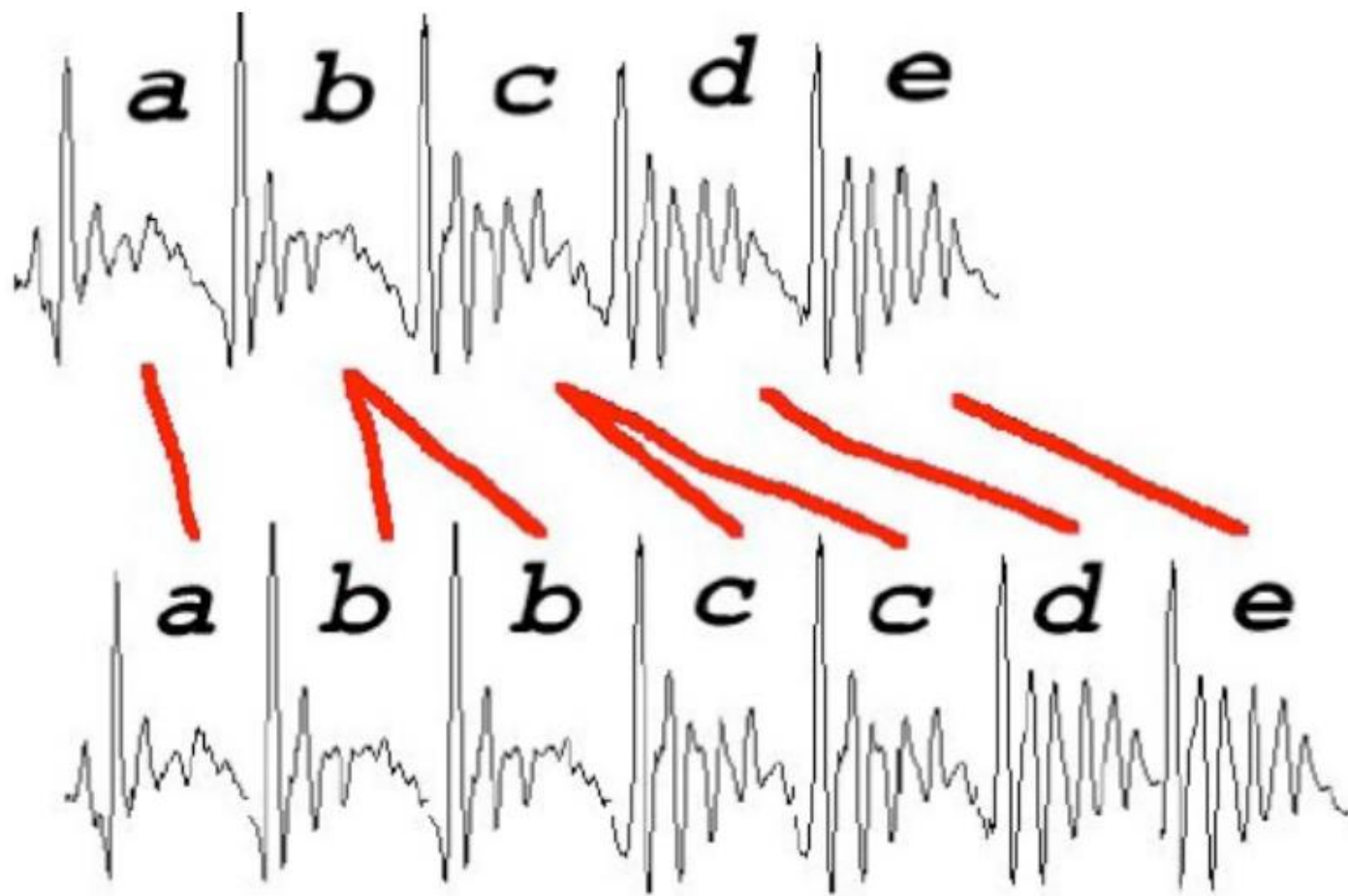
▶ 发展历程

- Single inventory: diphone synthesis [Moullnes; '90]
- Multiple inventory; unit selection synthesis (基元选择)
 - ATR v-Talk [Sagisaka; '92], CHATR [Black; '96]
 - AT&T Next-Gen TTS [Beutnagel; '99]

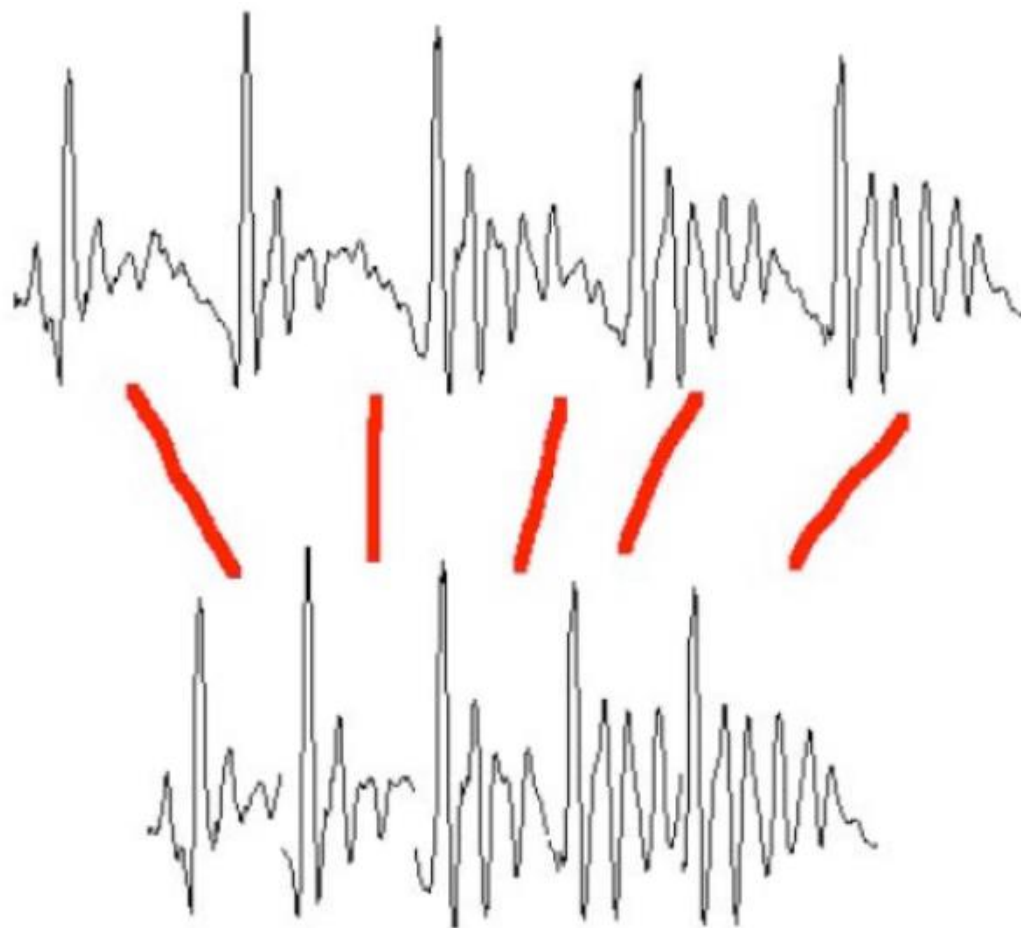
波形拼接系统框图



信号处理-复制/延长时长



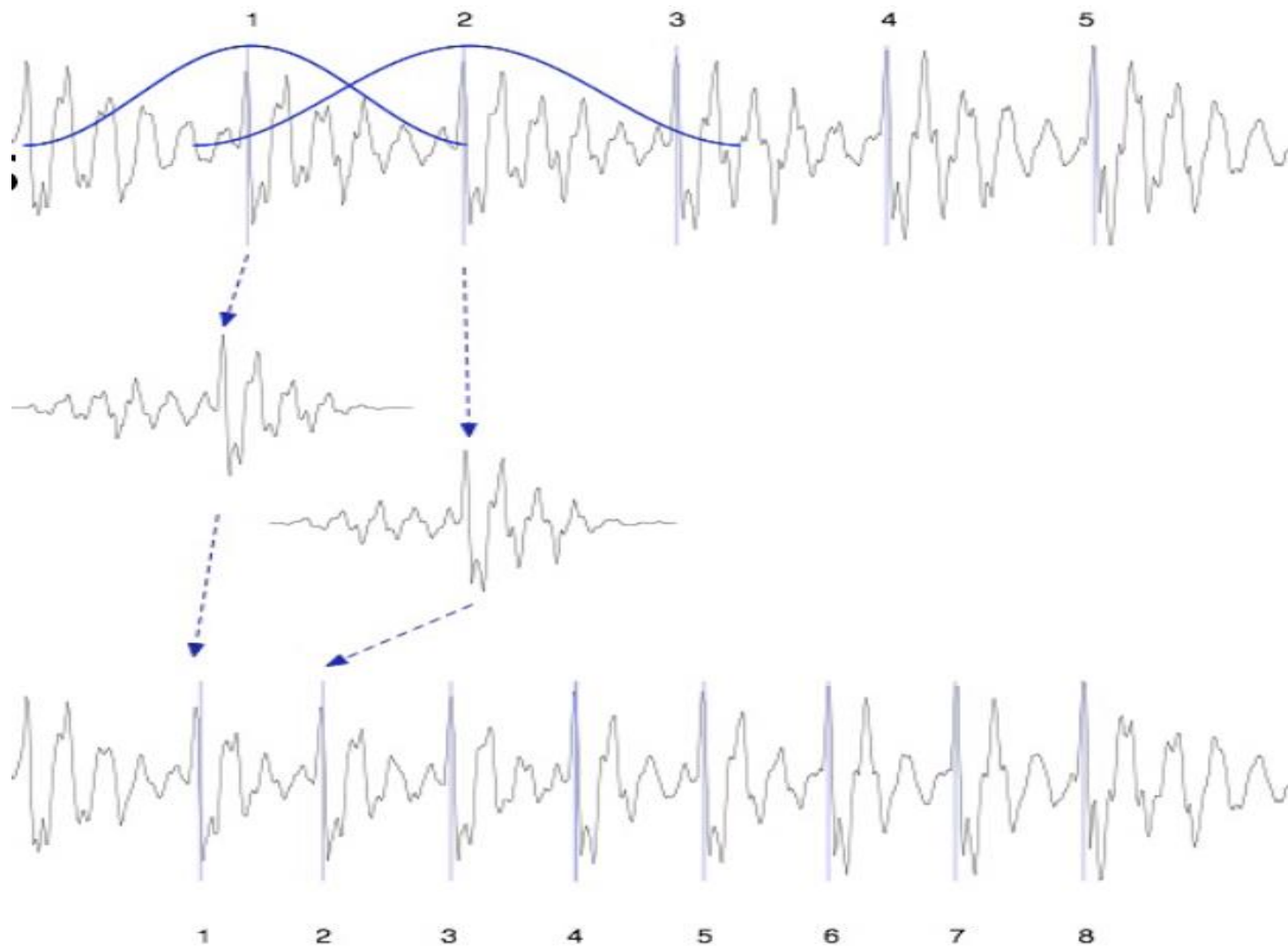
信号处理-提高/降低声调



PSOLA（基音同步叠加）合成

- 80年代末，由F. CharPentier等人提出的基音同步叠加技术PSOLA（Pitch Synchronous Overlap Add）是一种较好的波形拼接技术。
- 既能保持原始发音的主要音段特征，又能在拼接时灵活调节其音高和音长等韵律特征。
- 特点：根据语义对拼接单元的韵律特征进行调整，使合成波形既保持原始语音基元的主要音段特征，又使拼接单元的韵律特征符合语义，从而获得很高的可懂度和自然度。

PSOLA (基音同步叠加) 合成



波形合成法优缺点

▶ 优点

- 波形拼接方法简单直观，运算量小；
- 能方便地控制语音信号的韵律参数，能够合成自然连续语流。

▶ 缺点

- 音库往往非常庞大，不利于在嵌入式设备上实现。
- 在拼接时，两个相邻的声音单元之间谱的不连续，也容易造成合成音质的下降。

基于HMM的统计参数合成

▶ 技术实现

– 训练阶段

- 利用声码器提取训练语音的声学特征参数
- 训练统计声学模型

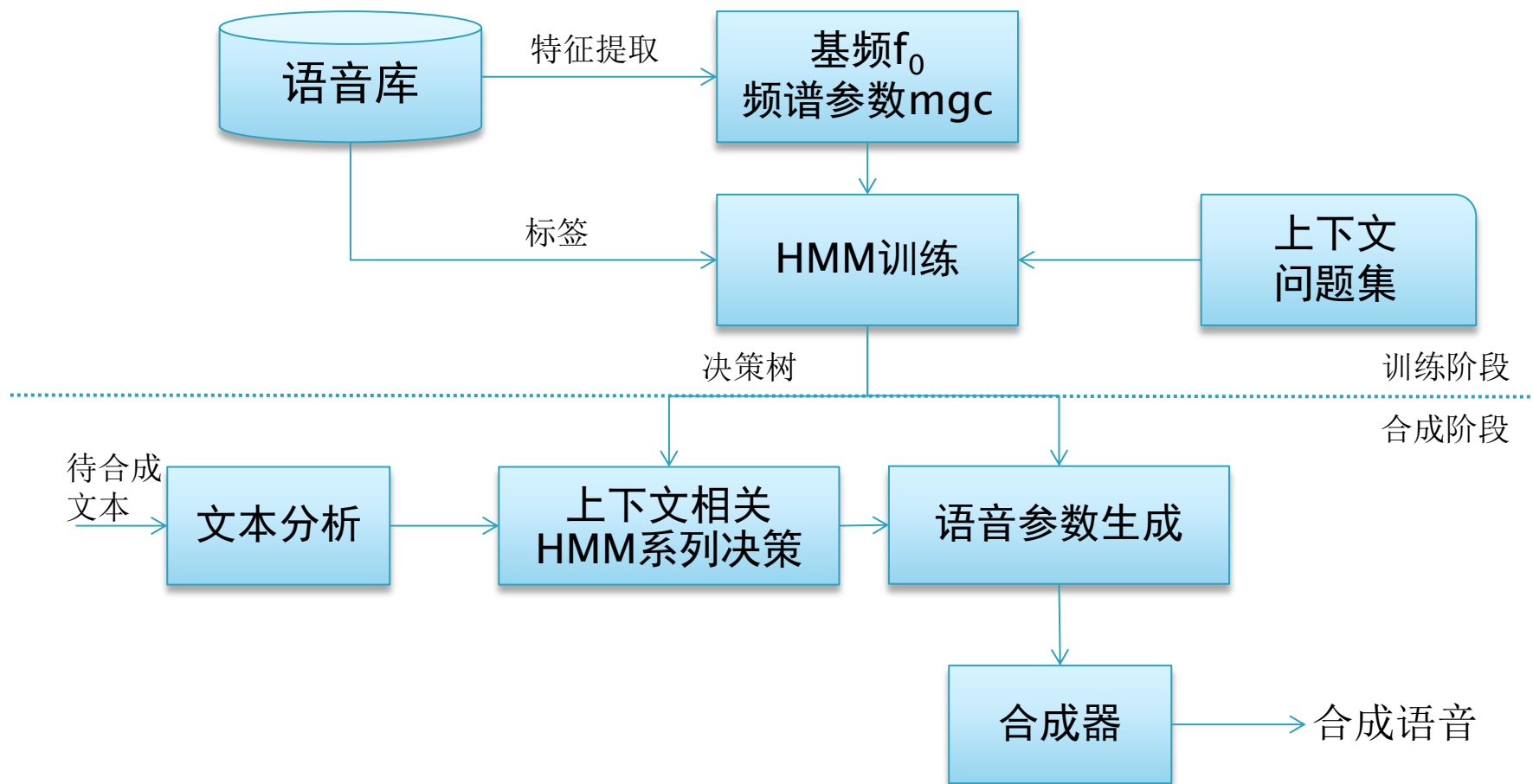
– 合成阶段

- 由前端文本分析得到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]

HMM语音合成系统框架

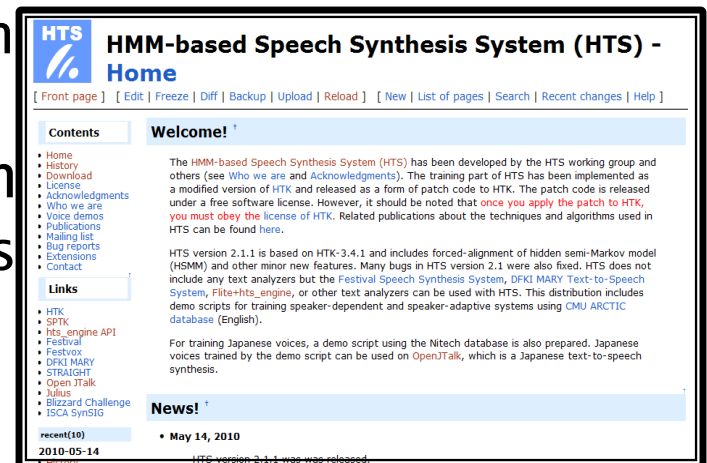


HMM优势

- ▶ HMM对声学参数的有效建模,自动快速的构建系统
 - ✓ 合成语音整体非常平稳,自然度高
 - ✓ 对说话人、表达方式甚至语种的依赖性非常小
 - ✓ HTS提供了一个便于开展工作的研究平台
 - ✓ STRAIGHT等高性能语音分析合成算法的出现为该方法走向实用提供了可能
 - ✓ 模型的灵活性给基于统计的参数合成方法提供了广阔的发展空间
 - 模型自适应,模型内插
 - 不同情感/风格的合成

HTS

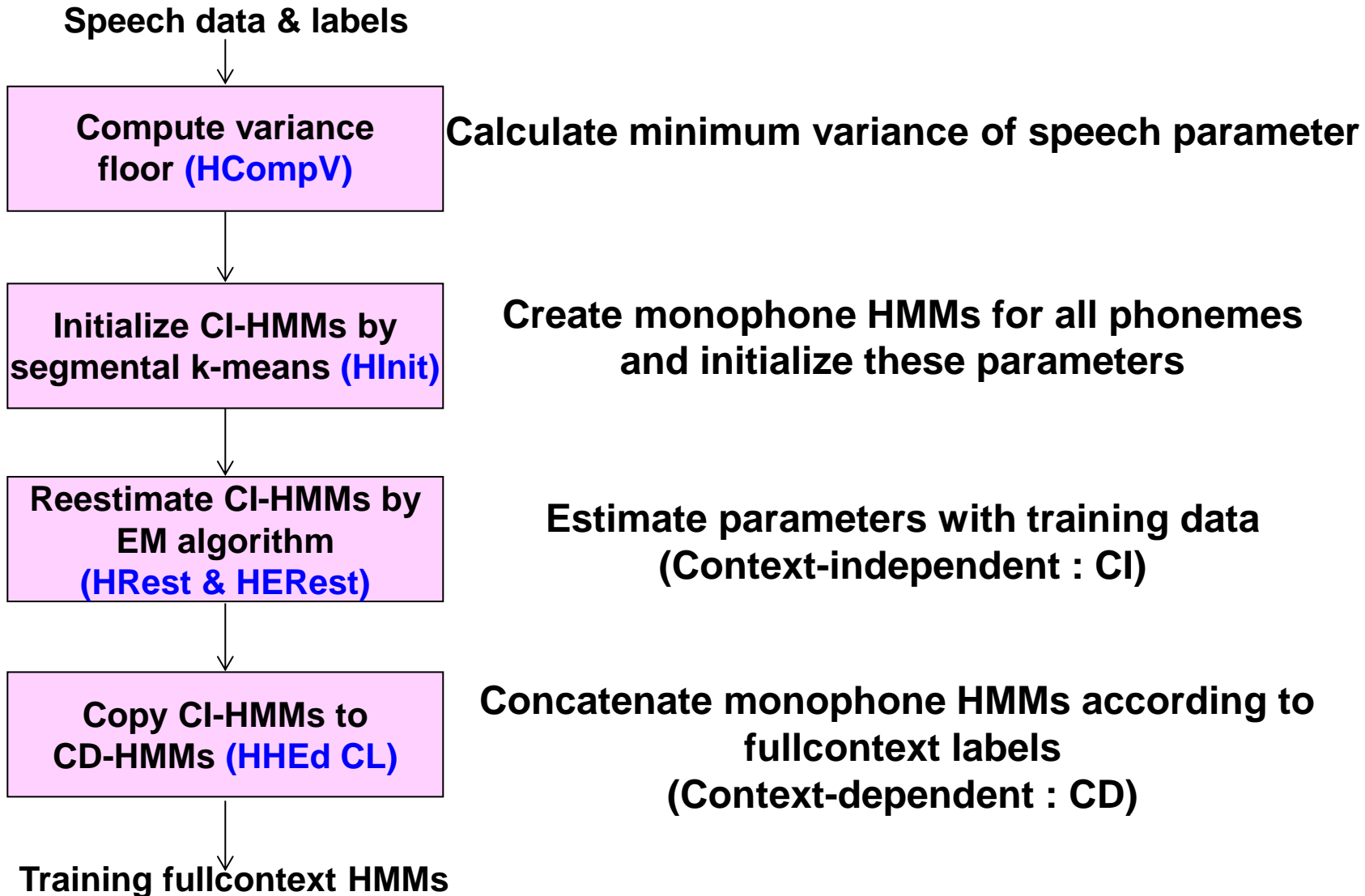
- ▶ HMM-based Speech Synthesis System (HTS)
 - Released as a form of patch code to HTK
 - Under the New and Simplified BSD license
 - HTS-users mailing list
 - Over 500 posts per year
 - All posts are archived & searchable
 - Bug reports, Q&A, announce
 - Becoming a research platform
 - Using by various organizations e. g., Microsoft, IBM



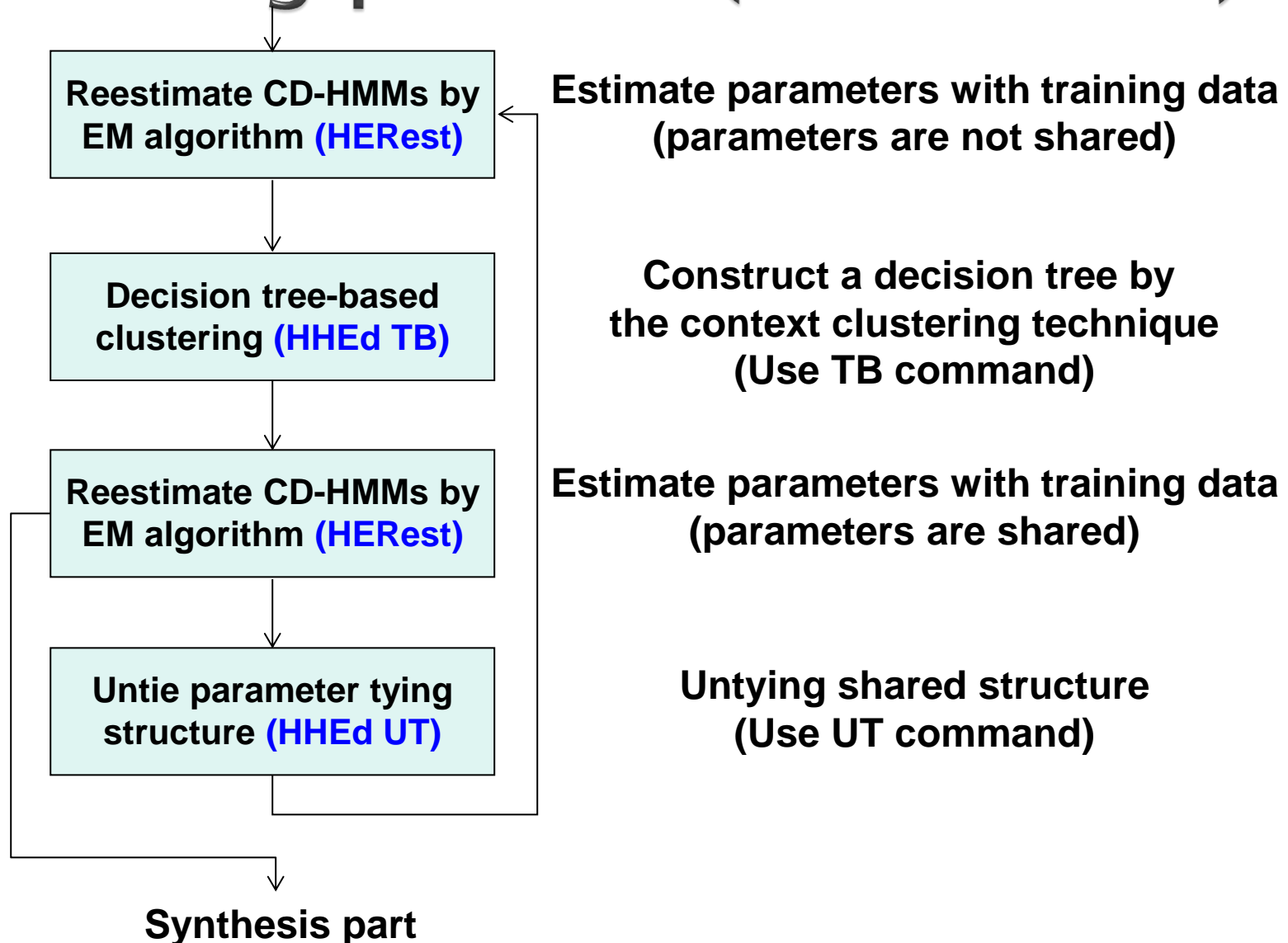
The screenshot shows the homepage of the HMM-based Speech Synthesis System (HTS). The page has a blue header with the HTS logo and the title "HMM-based Speech Synthesis System (HTS) - Home". Below the header, there are navigation links: "[Front page] [Edit] [Freeze] [Diff] [Backup] [Upload] [Reload] [New] [List of pages] [Search] [Recent changes] [Help]". The main content area is divided into three sections: "Contents", "Links", and "News!". The "Contents" section lists various links such as Home, History, Download, License, Acknowledgments, Who we are, Voice demos, Publications, Mailing list, Bug reports, Extensions, and Contact. The "Links" section lists external links like HTK, SPK, hts_engine API, Festival, Festvox, DRKI MARY, STRAIGHT, Open JTalk, Julius, Blizzard Challenge, and ISCA SynSIG. The "News!" section has a sub-header "recent(10)" and a date "2010-05-14" with a bullet point "• May 14, 2010". Below the news section, there is a small note: "HTS version 3.1.1 was released."

<http://hts.sp.nitech.ac.jp/>

Training process (monophone)



Training process (fullcontext)



HTS训练部分

- ▶ 将HMM的转移矩阵用一个时长(dur)模型替代；
- ▶ 用多空间概率分布对清浊音段进行联合建模。

HTS合成部分

1. 通过一定的语法规则、语言学的规律得到所需的上下文信息，标注在label文件中；
2. 待合成的label经过训练部分得到的决策树决策，得到语境最相近的叶节点HMM，做为模型的决策；
3. 由决策出来的模型算出合成的基频、频谱参数，然后根据时长模型得到各个状态的帧数，由基频、频谱参数的均值和方差算出在相应状态的持续时长帧数内的各维参数数值，结合动态特征，最终算出合成参数；
4. 由算出的参数构建源-滤波器模型，合成语音。

中文语音合成的实现过程

- ▶ 语料库准备
- ▶ 问题集的设计

语料库的实现

选取1000句具有中文语言学特点的文本语料，并找专业人士录制语音

语料文本设计

mono标签

单音素标签信息包含音素的两个边界时间，格式为：
“开始时间 终止时间 当前音素”

带有时间戳的包含上下文信息的标签格式为：“开始时间 终止时间 中文 标签格式”

full标签

语料库的实现

音频文件、标签文件构成了训练阶段所使用的语料库；
语料越多越好

用于训练的mono标签

```
1 0 9300000 sil
2 9300000 11100000 w
3 11100000 13200000 uo3
4 13200000 15000000 y
5 15000000 16200000 ve4
6 16200000 17100000 l
7 17100000 18000000 ai2
8 18000000 19500000 y
9 19500000 21000000 ve4
10 21000000 22800000 x
11 22800000 24000000 i3
12 24000000 24900000 h
13 24900000 26999970 uan1
14 26999970 28200000 ch
15 28200000 30600000 ih1
16 30600000 30900000 pau
17 30900000 31500000 zh
18 31500000 33000000 ong1
19 33000000 33900000 g
20 33900000 35400000 uo2
21 35400000 36600000 c
22 36600000 40199970 ai4
23 40199970 45900000 sil
24
```

上下文信息

中文标签格式:

$p1 \wedge p2 - p3 + p4 = p5 / A : a1 _ a2 _ a3 / B : b1 _ b2 / C : c1 _ c2 / D : d1 _ d2 _ d3 / E : e / F : f$

表 3-3 中文词性的分类与标识符

词性	形容词	连词	副词	感叹词	方位词	前缀词
标识符	a	c	d	e	f	h
词性	成语	缩略词	后缀词	数词	拟声词	介词
标识符	i	j	k	m	o	p
词性	量词	代词	处所词	时间词	虚词	动词
标识符	q	r	s	t	u	v
词性	标点	字符串	连接词	状态词		
标识符	w	x	y	z		

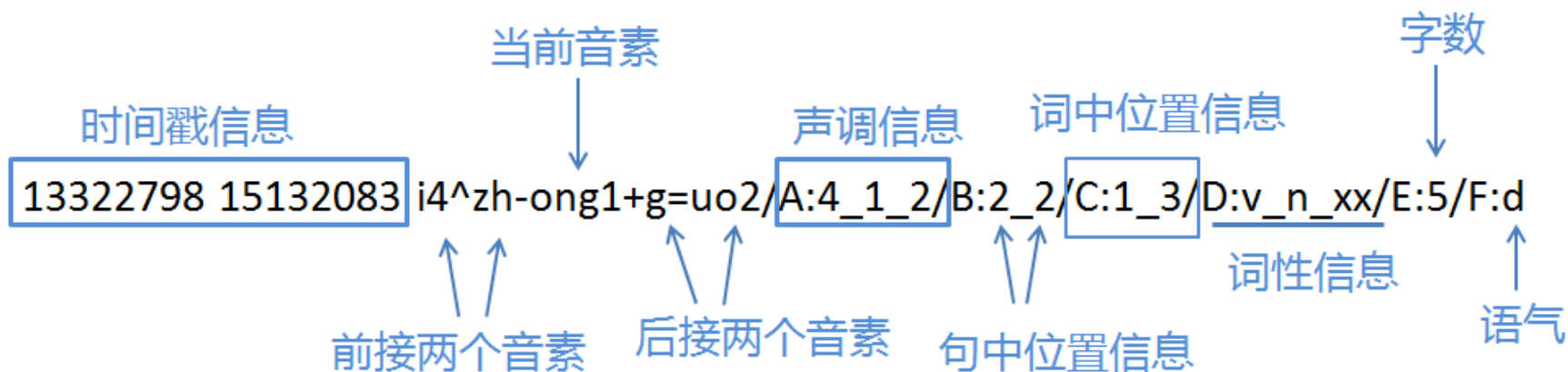
表 3-1 声调类型及其代表数字

声调类型	第一声阴平	第二声阳平	第三声上声	第四声去声	零声调
代表数字	1	2	3	4	0

表 3-4 句子语气类型与标识符

语气类型	陈述句	感叹句	祈使句	疑问句
标识符	d	e	i	q

上下文信息



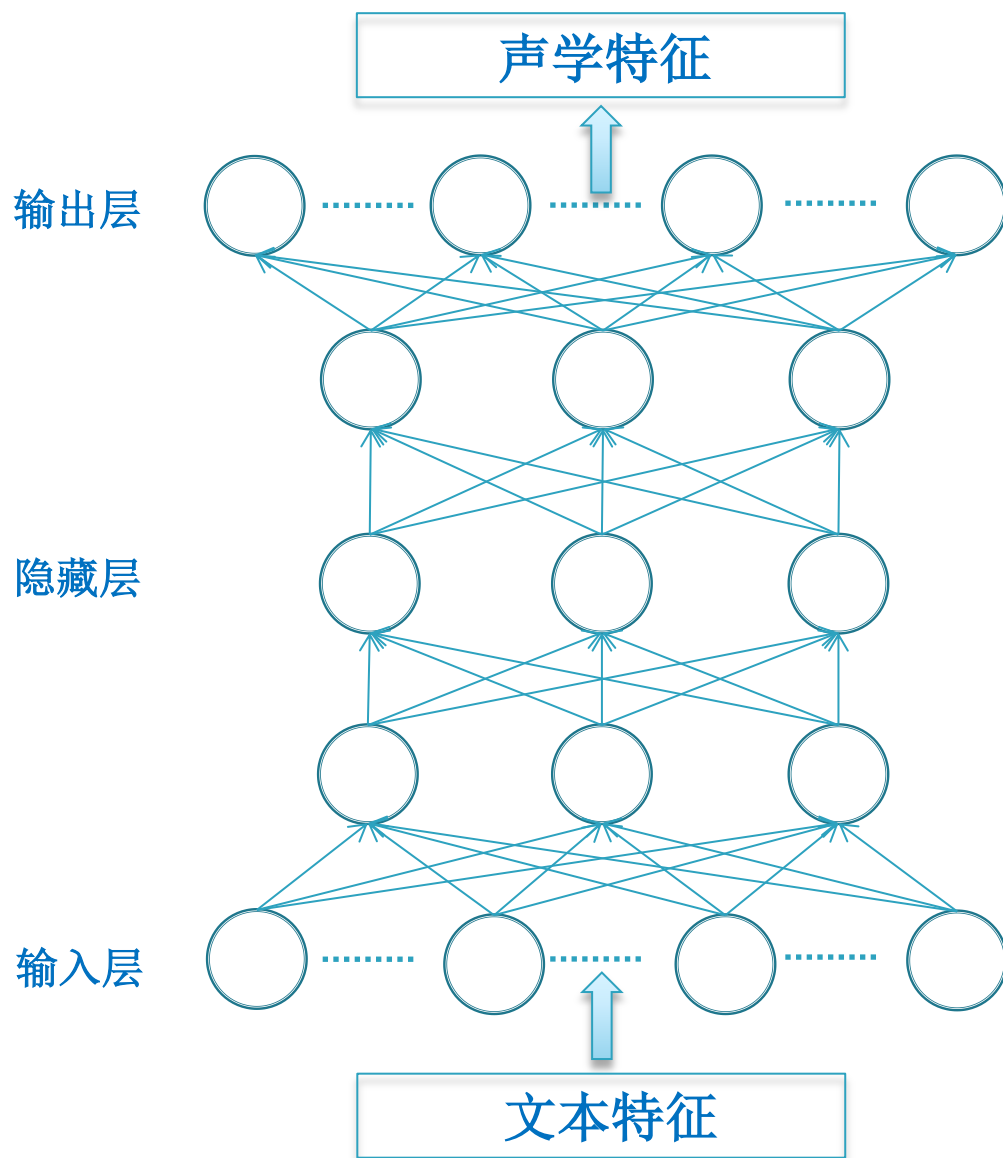
用于训练的full标签

```
1 0 7002364 xx^xx-sil+w=o3/A:xx_xx_xx/B:xx_xx/C:xx_xx/D:xx_xx_xx/E:5/F:d
2 7002364 7919068 xx^sil-w+o3=sh/A:xx_3_4/B:0_4/C:1_1/D:xx_r_v/E:5/F:d
3 7919068 8908143 sil^w-o3+sh=i4/A:xx_3_4/B:0_4/C:1_1/D:xx_r_v/E:5/F:d
4 8908143 10596809 w^o3-sh+i4=zh/A:3_4_1/B:1_3/C:1_1/D:r_v_n/E:5/F:d
5 10596809 12406094 o3^sh-i4+zh=ong1/A:3_4_1/B:1_3/C:1_1/D:r_v_n/E:5/F:d
6 12406094 13322798 sh^i4-zh+ong1=g/A:4_1_2/B:2_2/C:1_3/D:v_n_xx/E:5/F:d
7 13322798 15132083 i4^zh-ong1+g=uo2/A:4_1_2/B:2_2/C:1_3/D:v_n_xx/E:5/F:d
8 15132083 15904045 zh^ong1-g+uo2=r/A:1_2_2/B:3_1/C:2_3/D:v_n_xx/E:5/F:d
9 15904045 17134358 ong1^g-uo2+r=en2/A:1_2_2/B:3_1/C:2_3/D:v_n_xx/E:5/F:d
10 17134358 18268177 g^uo2-r+en2=sil/A:2_2_xx/B:4_0/C:3_3/D:v_n_xx/E:5/F:d
11 18268177 20559938 uo2^r-en2+sil=xx/A:2_2_xx/B:4_0/C:3_3/D:v_n_xx/E:5/F:d
12 20559938 25077551 r^en2-sil+xx=xx/A:2_xx_xx/B:xx_xx/C:xx_xx/D:n_xx_xx/E:5/F:d
```

问题集的设计

Vowel 元音 (韵母) a o e i u v ai ei ao ou an en in ang eng ing ong un ia ie iao ian iang iong ua uai uan ue uang uo 30个
Diphthong 双元音 ia ie iao ian iang iong ua uai uan ue uang uo 12个
TypeA 含有a的韵母 a ai ao an ang ia iao ian iang ua uai uan uang 13个
TypeI 含有i的韵母 i ai ei in ing ia ie iao ian iang iong uai 12个
TypeE 含有e的韵母 e ei eng ie ue 5个
TypeU 含有u的韵母 u ou un ua uai uan ue uang uo 9个
TypeO 含有o的韵母 o ao iao ong iong uo 6个
TypeV 含有v的韵母 v
Anterior_Nasal_Vowel 前鼻韵母 an en in un
Posterior_Nasal_Vowel 后鼻韵母 ang eng ing ong
Consonant 辅音 (声母)
Stop 塞音
Aspirated_Stop 送气塞音
Non_Aspirated_Stop 不送气塞音
Fricative 擦音
Breath_Fricative 清擦音
Voiced_Fricative 浊擦音
Affricate 塞擦音
Aspirated_Affricate 送气塞擦音
Non_Aspirated_Affricate 不送气塞擦音
Labial 唇音
Lateral 边音
Front 舌尖前音
Central 舌尖中音
Back 舌尖后音
Surface 舌面音
Root 舌根音
Pos==a 是否为形容词 (共25个)
Toner=1 是否为一声调 (共5个)

DNN合成法



DNN合成法

▶ 输入文本特征

- 逐帧输入
- 根据问题集生成，维度与问题数一致



▶ 输出声学特征

- 逐帧输出
- 对应声学特征(MGC:35, logF0:1, voiced/unvoiced:1)
- 输出维度: $3 \times (35 + 1) + 1 = 109$

深度生成模型

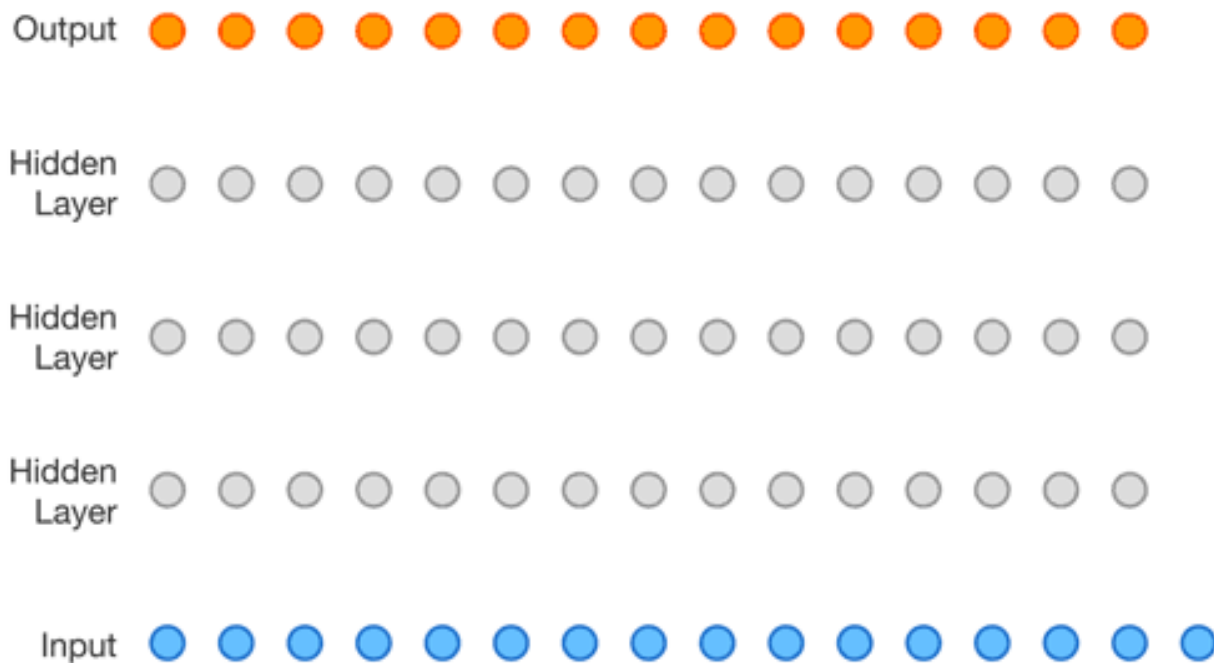
- ▶ 参数模型

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^T p(x_t|x_1, \dots, x_{t-1}, \mathbf{h})$$

- ▶ 其中 \mathbf{h} 代表输入的参数，包括想要说的文字，以及要模仿的口音信息。
- ▶ 为了处理超长距离的依赖关系，可采用因果卷积网络。

因果卷积网络

- ▶ 可以在时序上展开的多层卷积神经网络，而每一步当前所输出的结果，只能够依赖于过去的输出，不能够依赖于未来的输出。



WaveNet合成效果(MOS)

- ▶ WaveNet—Deep Mind研发的深度合成系统。
- ▶ MOS最大值5，越高越好。

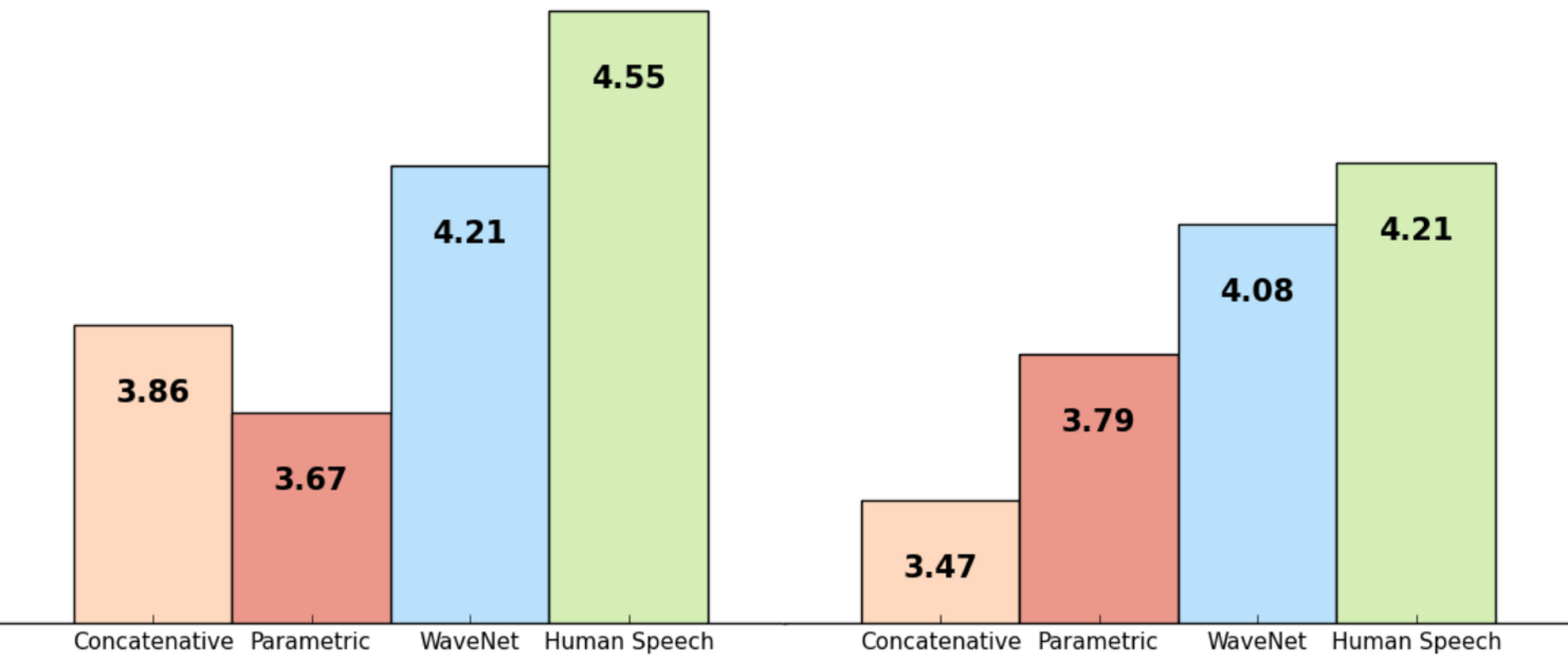
Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	4.21 ± 0.081	4.08 ± 0.085
Natural (8-bit μ -law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

From: A ãron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, Koray Kavukcuoglu: WaveNet: A Generative Model for Raw Audio. CoRR abs/1609.03499 (2016)

WaveNet合成效果(MOS)

US English

Mandarin Chinese



From: <https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

语音合成发展方向

- ▶ 提高语音合成的表现力
 - 情感化合成
- ▶ 多语种合成系统研究
- ▶ 方言合成（闽南语、维吾尔语...）
- ▶ 个性化语音合成
 - 话者转换（Voice Conversion）：对特定发音人的模仿
 - 话音变换（Voice Morphing）
 - 歌唱合成（Singing TTS）

Thank you!

Any questions?