深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

语音合成：从入门到精通

第三讲：传统语音合成算法

主讲人 陈云琳

出门问问资深语音工程师
毕业于西北工业大学ASLP

目录

目录

语音合成系统框架

输入文本

西北工业大学是唯一 一所以同时发展三航为特色的985好大学。

语言学模型与规则

前端 文本分析 Text Analysis

- 文本正则化/Text Normalization(TN)
- 分词/Word Segmentation
- 词性预测/Part-of-speech
- 注音(多音字)/Polyphone
- 韵律预测/Prosody

西北工业大学是唯一一所以同时发展三航为特色的九八五好大学。

西北/工业/大学/是/唯一/一所/以/同时/发展/三航/为/特色/的/九八五/好/大学。

西北/n工业/n大学/n是/v唯一/a一所/a以/v同时/a发展/v三航/n为/v特色/a的/n九八五/n好/a大学/n。

xi1 bei3 gong1 ye4 da4 xue2 shi4 wei2 yi1 yi1 suo3 yi3 tong2 shi2 fa1 zhan3 san1 hang2 wei2 te4 se4 de5 jiu3 ba1 wu3 hao3 da4 xue2 pau0

西北#2工业#1大学#2是#2唯一#1一所#2以#1同时#1发展#2三航为#1特色的#2九八五#1好大学。

后端 语音合成 Acoustic Generation

波形拼接合成

统计参数合成

- 帧级建模
  - 时长模型
  - 声学模型：文本分析结果 → 声学特征
- 声码器：声学特征 → 语音波形采样点

输出声音

输出声音

回顾 – 文本分析

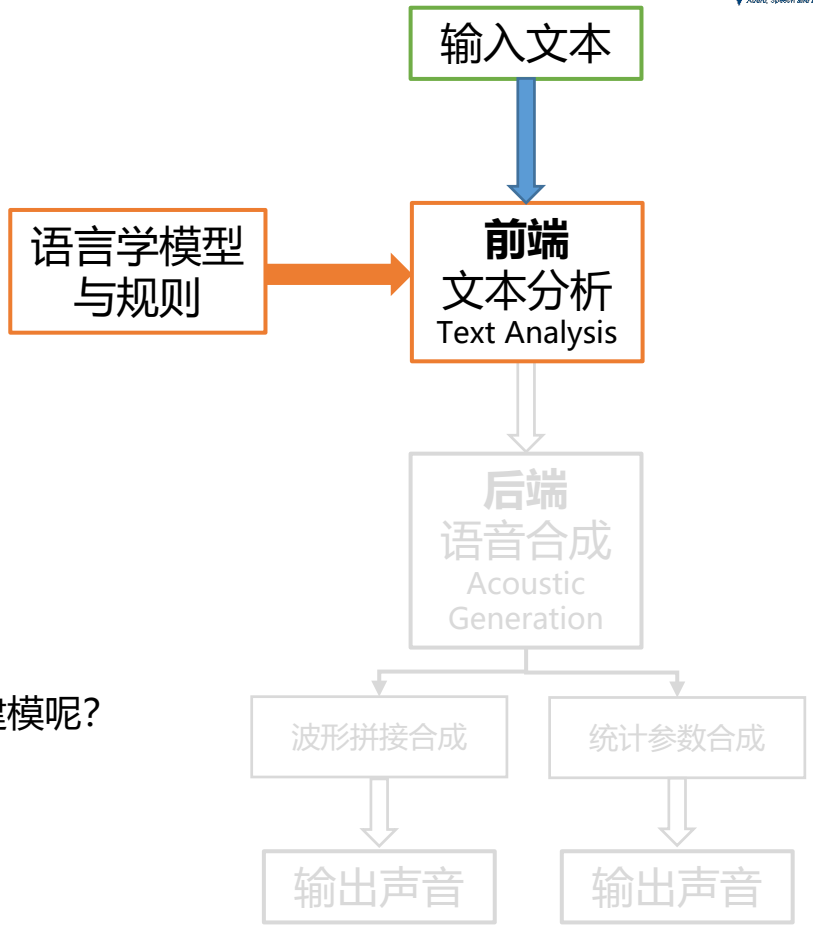
文本分析的基本组成

- 文本正则化/分词/词性/g2p/韵律

文本分析各个模块的方法

- 文本正则化：基于规则的方法
- 分词：字典/CRF/BLSTM+CRF/BERT
- 注音：N-Gram/CRF/BLSTM/seq2seq
- 韵律：CRF/BLSTM+CRF/BERT

如何把这些模块组合在一起，应用到语音合成建模呢？



回顾 - 语音合成概率公式

$$\hat{\mathcal{O}} = \arg\max_{\mathcal{O}} p(\mathcal{X} \mid \mathcal{O})$$ Vocoder analysis

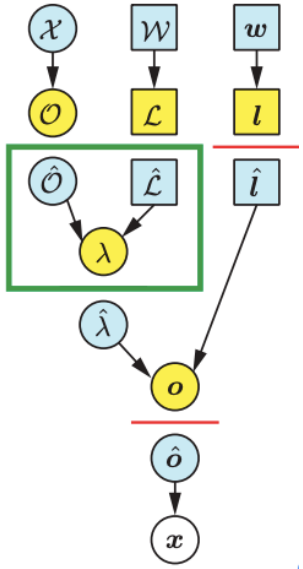
$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid \mathcal{W})$$ Text analysis

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\mathcal{O}} \mid \hat{\mathcal{L}}, \lambda) p(\lambda)$$ ***Extract rules***

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l} \mid \boldsymbol{w})$$ Text analysis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ ***Parameter generation***

$$\bar{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(\hat{\boldsymbol{o}}) = p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$ Vocoder synthesis



2017 - Heiga et al. - Generative Model-Based Text-to-Speech Synthesis （上文公式来源）

语音合成 – 后端模型

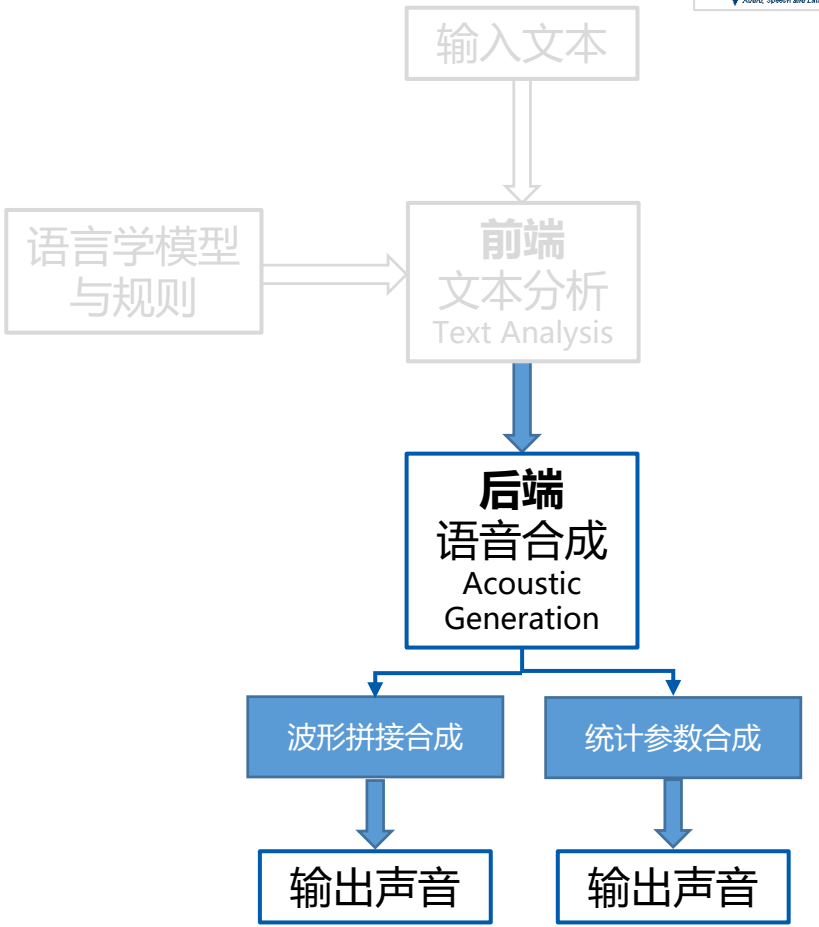
输入：文本分析结果

输出：语音波形采样点

参数语音合成方法

- 声学模型(Acoustic Model)
  - 基于HMM的统计参数语音合成
  - 基于NN的统计参数语音合成
- 声码器(Vocoder)
  - 基于源-滤波器(Source-Filter)的声码器
  - 基于NN的声码器(第6节课)

单元拼接语音合成方法



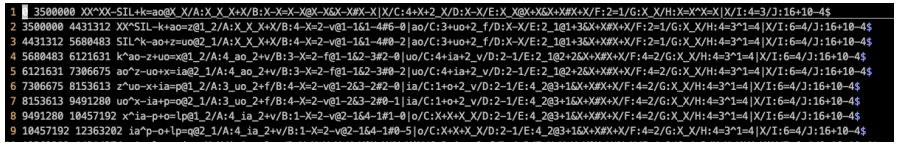
目录

2. 基于隐马尔可夫(HMM)的统计参数语音合成

音频语音与语言处理研究组 SLP
Audio, Speech and Language Processing Group NPU

- 2.1 统计参数语音合成框架
- 2.2 隐马尔可夫模型(HMM)
- 2.3 多空间概率分布MSD-HMM
- 2.4 基于HMM的参数语音合成

参数语音合成方法系统框架



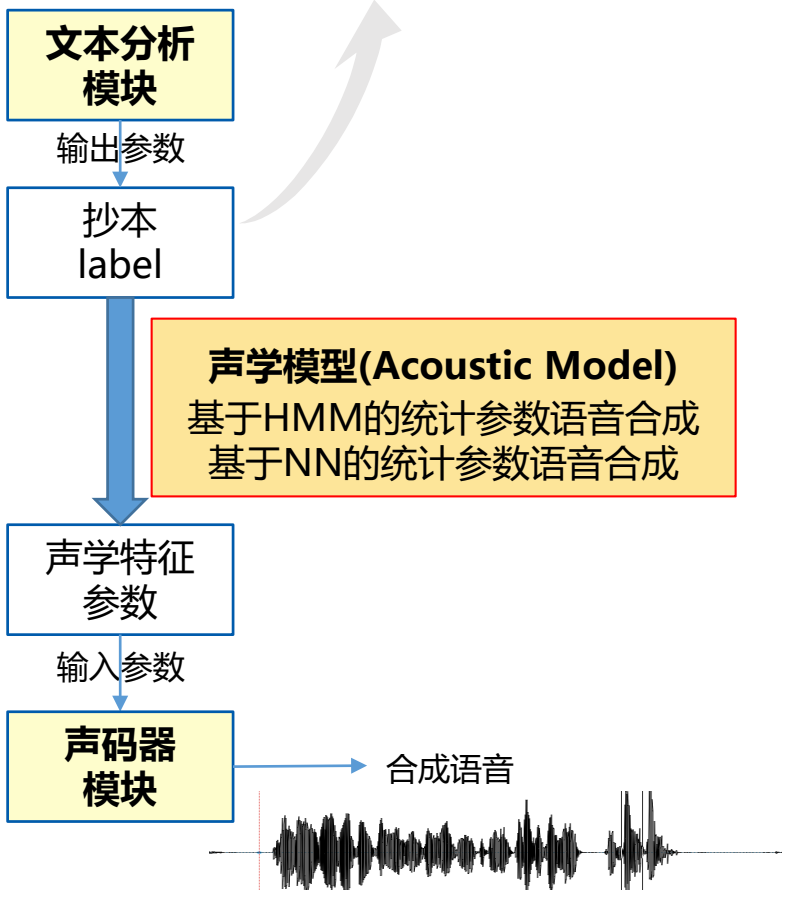
声学模型生成（Acoustic Generation）

建立**文本分析**和**声码器**之间的桥梁，

结合声码器最终合成语音。

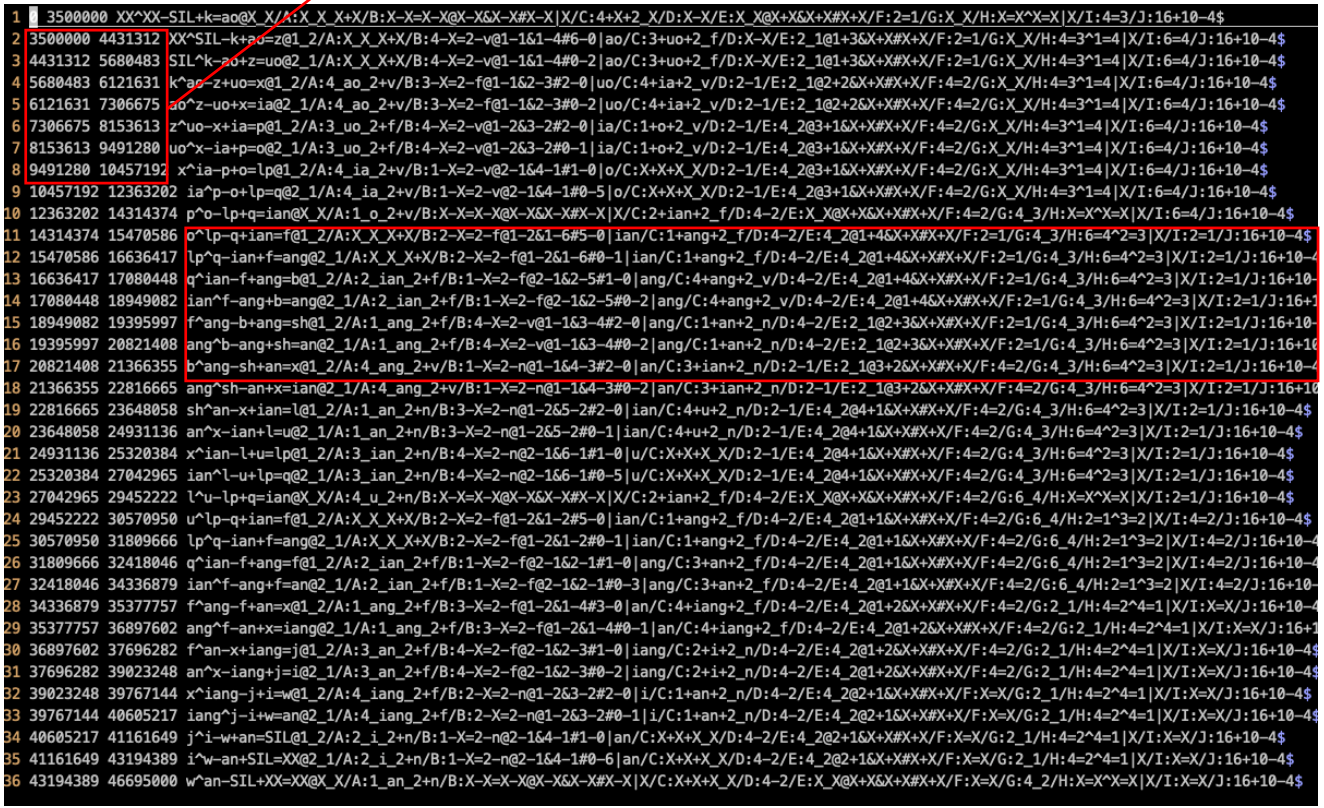
参数语音合成方法

- 声学模型(Acoustic Model)
  - 基于HMM的统计参数语音合成
  - 基于NN的统计参数语音合成
- 声码器(Vocoder)
  - 基于源-滤波器(Source-Filter)的声码器
  - 基于NN的声码器(第6节课)



输入：抄本

声学模型使用标注的时长拓展为帧级别信息



各种语言学信息：

音素、音节、词等的位置，词性、声调、重读、韵律等特征

用于声学模型训练和解码

p1^p2-p3+p4=p5@p6_p7/A:a1_a2&a3_a4/B:b1_b2#b3_b4！b5_b6/C:c1+c2/D:d1_d2/E:e1+e2/F:f1_f2/G:g1+g2+g3/H:

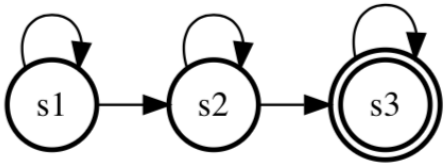
音素(Phone)

- **词** 并不是一个语言的基本发音单元，以词为建模单元无法**共享**这些发音的基本单元。
- 发音的基本单元: **音素**
- 静音Silence(SIL)

通过词典映射

one	W AA N
two	T UW
three	TH R IY
……	

单音素HMM拓扑结构



- **英文音素（CMU phone，39)**

AA AE AH AO AW AX AXR AY
B BD CH D DD DH DX EH ER EY
F G GD HH IH IX IY JH K
KD L M N NG OW OY P PD
R S SH T TD TH TS UH UW
V W X Y Z ZH

- **中文音素（可以认为声韵母就是音素)**

a o e i u v
b p m f d t n l g k h j q x
zh ch sh z c s y w
ai ei ui ao ou iu ie ue er
an en in un vn
ang eng ing ong

上下文属性集

p1^p2-p3+p4=p5@p6_p7/A:a1_a2&a3_a4/B:b1_b2#b3_b4！b5_b6/C:c1+c2/D:d1_d2/E:e1+e2/F:f1_f2/G:g1+g2+g3/H:

p1：LL phoneme, 当前音素的左边的左边的音素
p2：L(eft) phoneme, 当前音素的左边的音素
p3：C(urrent) phoneme, 当前音素
p4：R(ight) phoneme, 当前音素右边的音素
p5：RR phoneme, 当前音素的右边的右边的音素
p6：从左往右数，当前音素在当前音节中的位置
p7：从右往左数，当前音素在当前音节中的位置
a1：当前音节的前一个音节是否要重读
a2：当前音节的前一个音节的音素数目
a3：从左往右数，当前音节在韵律词中的位置
a4：从右往左数，当前音节在韵律词中的位置
b1：当前音节是否要重读
b2：当前音节拥有的音素数目
b3：从左往右数，当前音节在当前词中的位置
b4：从右往左数，当前音节在当前词中的位置

b5：从上一个重读音节到当前音节之间的音节数目
b6：从当前音节到下一个重读音节之间的音节数目
c1：当前音节的下一个音节是否要重读
c2：下个音节拥有的音素数目
d1：当前词的前一个词的词性
d2：当前词的前一个词拥有的音节数目
e1：当前词的词性
e2：当前词拥有的音节数目
f1：当前词的下一个词的词性
f2：当前词的下一个词拥有的音节数目
g1：当前音节的前一个音节的声调
g2：当前音节的音调
g3：当前音节的下一个音节的声调

总结：音素、音节、词等的位置，词性、声调、重读、韵律等特征

输出： Acoustic Features(声学特征)

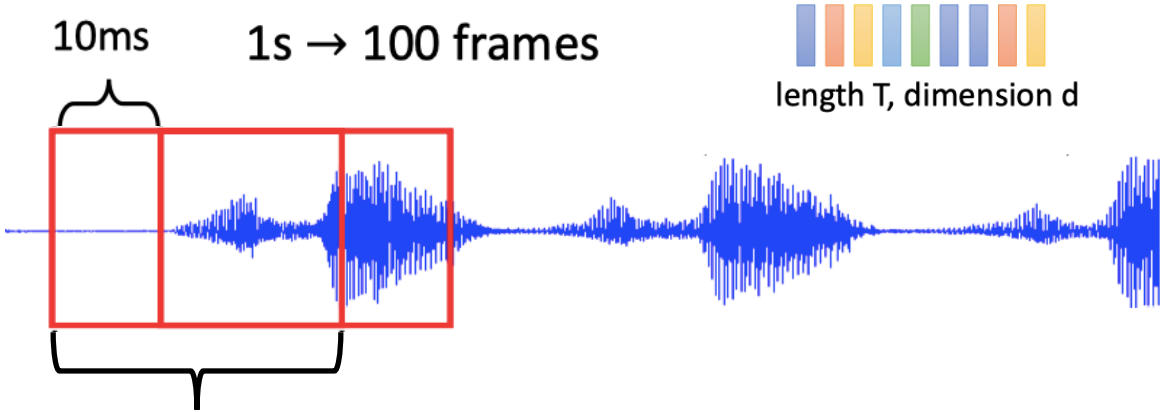
声学特征参数

参数类型	分类	提取方法（工具）
谱参数	MGC (mel-generalized cepstral)：梅尔生成倒谱 MCEP (mel-cepstrum)：梅尔倒谱 LSP (line spectrum pairs)：线谱对	Straight World SPTK etc.
激励参数： 基频F0(pitch)	清音(unvoiced)：b, p, t, g … 浊音(voiced)：a, o, e, i …	Reaper Straight World praat etc.

合成语音：MGC or MCEP or LSP + F0 ➔ 音频

Acoustic Features(声学特征)

声学特征：谱参数



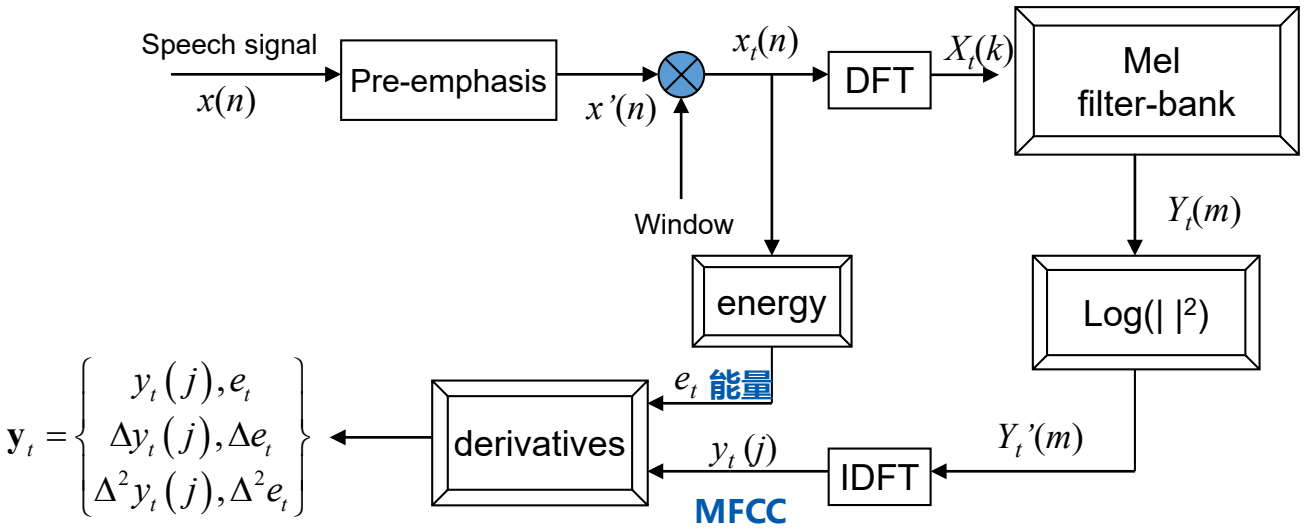
frame(帧)

- 400 sample points(16kHz)
- 41-dim mgc/lsp
- 80-dim filter bank(mel)
- 39-dim mfcc

...

Acoustic Features(声学特征)

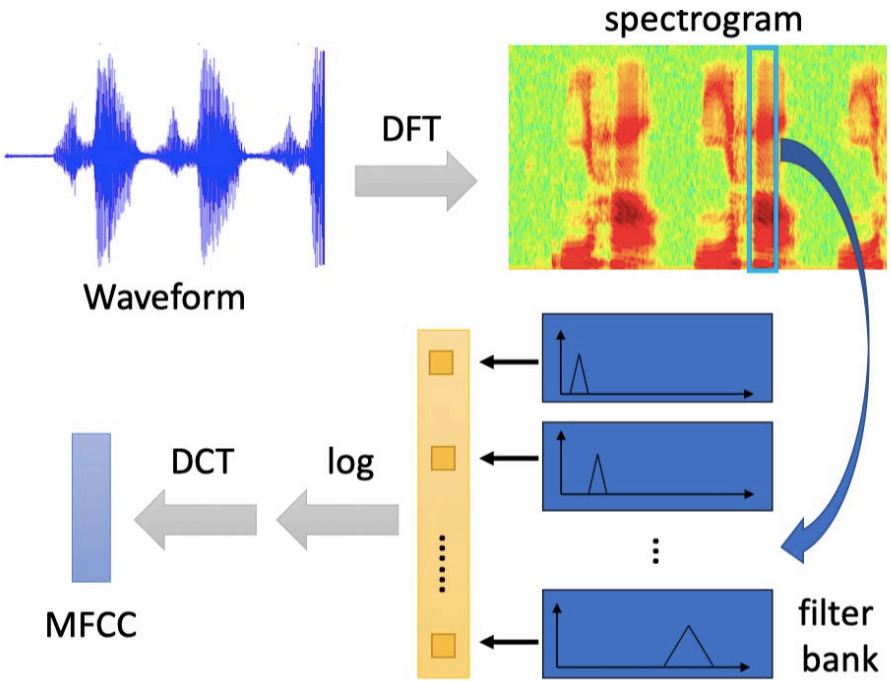
MFCC (Mel-Frequency Cepstral Coefficients)特征提取过程



数字语音处理：http://ocw.aca.ntu.edu.tw/ntu-ocw/ocw/cou/104S204/7

Acoustic Features(声学特征)

MFCC (Mel-Frequency Cepstral Coefficients)特征提取过程



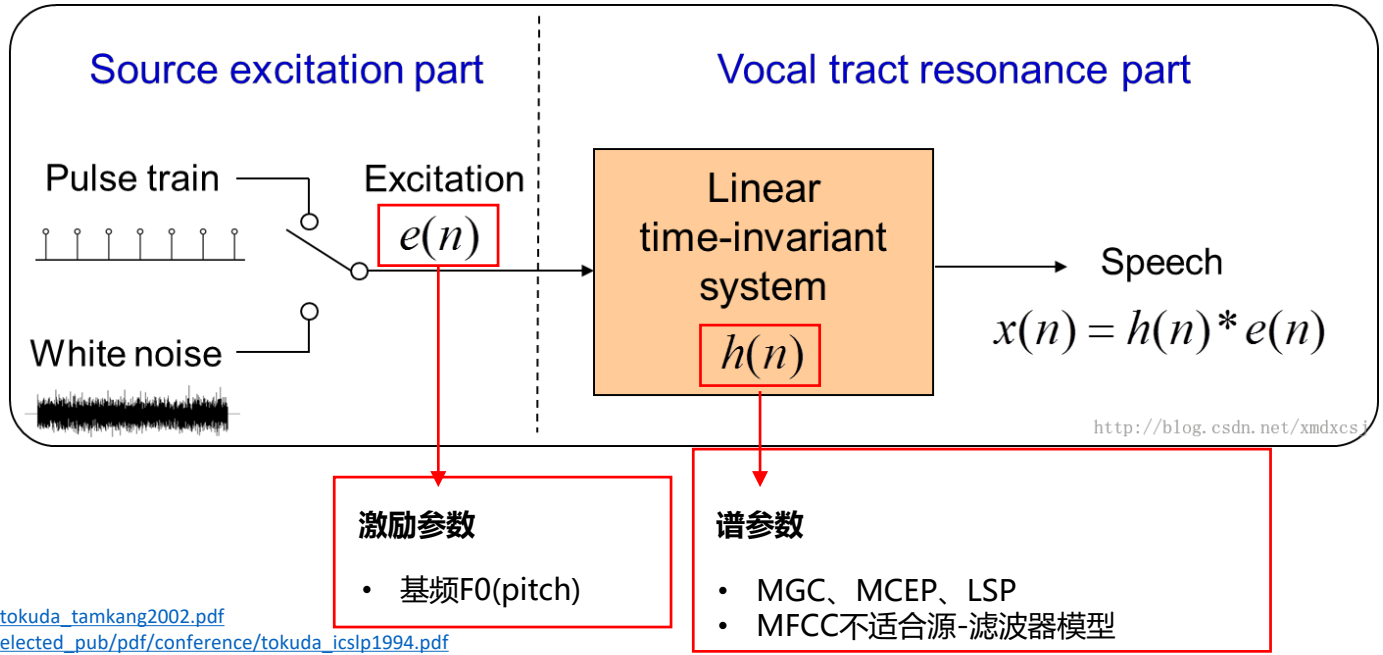
数字语音处理：http://ocw.aca.ntu.edu.tw/ntu-ocw/ocw/cou/104S204/7

Acoustic Features(声学特征)

- MGC、MCEP、LSP和MFCC什么区别呢?
  - 语音合成的声码器：使用源-滤波器模型
  - MFCC不适合源-滤波器模型，其他几个都可以，但是对合成质量各有不同

声码器：源-滤波器模型

激励部分　　声道模型（滤波器）



https://www.sp.nitech.ac.jp/~tokuda/tokuda_tamkang2002.pdf
http://www.sp.nitech.ac.jp/~tokuda/selected_pub/pdf/conference/tokuda_icslp1994.pdf

Acoustic Features(声学特征)

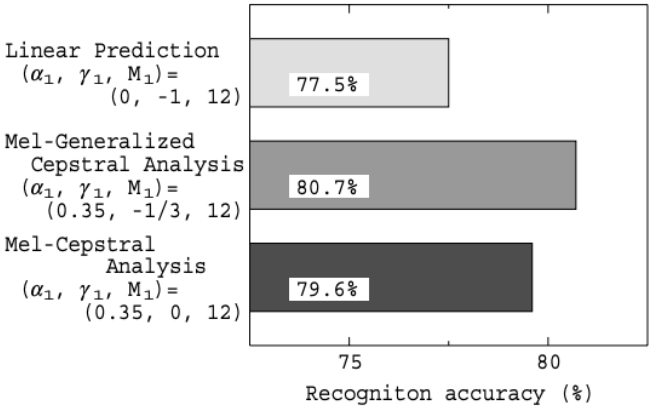
MGC、MCEP、LSP与MFCC有什么区别呢?

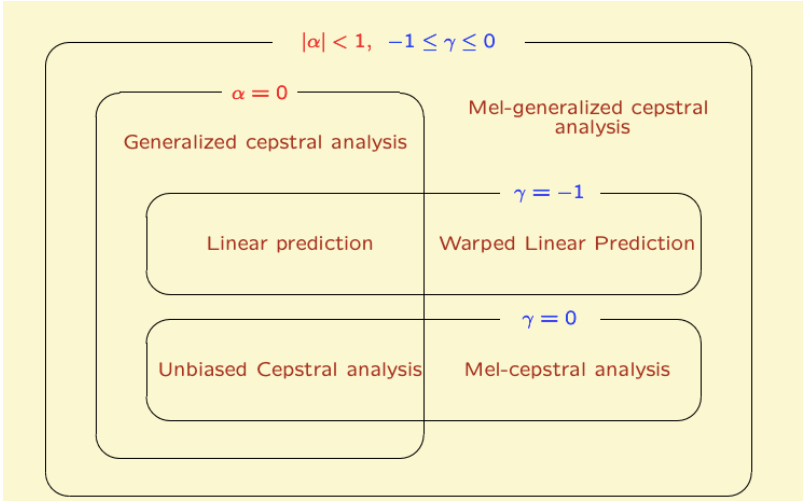
Mel-generalized cepstrum: $c_{\alpha,\gamma}(m)$

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}\right)$$

$$= \begin{cases} \left(1+\gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}\right)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}, & \gamma = 0 \end{cases}$$

$$z_\alpha^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}$$





https://www.sp.nitech.ac.jp/~tokuda/tokuda_tamkang2002.pdf
http://www.sp.nitech.ac.jp/~tokuda/selected_pub/pdf/conference/tokuda_icslp1994.pdf

Acoustic Features(声学特征)

MGC/LSP提取过程

```
    $(X2X) +sf $${raw} | $(PITCH) -H $(UPPERF0) -L $(LOWERF0) -p $(FRAMESHIFT) -s $${SAMPKHZ} -o 2 > lf0/$${base}.lf0; \
    if [ $(GAMMA) -eq 0 ]; then \
        $(X2X) +sf $${raw} | \
        $(FRAME) -l $(FRAMELEN) -p $(FRAMESHIFT) | \
        $(WINDOW) -l $(FRAMELEN) -L $(FFTLEN) -w $(WINDOWTYPE) -n $(NORMALIZE) | \
        $(MGCEP) -a $(FREQWARP) -m $(MGCORDER) -l $(FFTLEN) -e 1.0E-08 > mgc/$${base}.mgc; \
    else \
        if [ $(LNGAIN) -eq 1 ]; then \
            GAINOPT="-L"; \
        fi; \
        $(X2X) +sf $${raw} | \
        $(FRAME) -l $(FRAMELEN) -p $(FRAMESHIFT) | \
        $(WINDOW) -l $(FRAMELEN) -L $(FFTLEN) -w $(WINDOWTYPE) -n $(NORMALIZE) | \
        $(MGCEP) -a $(FREQWARP) -c $(GAMMA) -m $(MGCORDER) -l $(FFTLEN) -e 1.0E-08 -o 4 | \
        $(LPC2LSP) -m $(MGCORDER) -s $${SAMPKHZ} $${GAINOPT} -n $(FFTLEN) -p 8 -d 1.0E-08 > mgc/$${base}.mgc; \
    fi; \
```

工具

- HTS pipeline(http://hts.sp.nitech.ac.jp/) / SPTK
- Pysptk: https://github.com/r9y9/pysptk

2. 基于隐马尔可夫(HMM)的统计参数语音合成

例子

按照以下方式从盒子里抽球：开始时，从第一个盒子抽球的概率是0.2，从第二个盒子抽球的概率是0.4，从第三个盒子抽球的概率是0.4。以这个概率抽一次球后，将球放回；然后，从当前盒子转移到下一个盒子进行抽球，这一步的规则是，如果当前抽球的盒子是第一个盒子，则以0.5的概率仍然留在第一个盒子继续抽球，以0.2的概率去第二个盒子抽球，以0.3的概率去第三个盒子抽球。如果当前抽球的盒子是第二个盒子，则以0.5的概率仍然留在第二个盒子继续抽球，以0.3的概率去第一个盒子抽球，以0.2的概率去第三个盒子抽球。如果当前抽球的盒子是第三个盒子，则以0.5的概率仍然留在第三个盒子继续抽球，以0.2的概率去第一个盒子抽球，以0.3的概率去第二个盒子抽球。如此下去，直到重复三次，得到一个球的颜色的观测序列。

盒子	1	2	3
红球数	5	4	7
白球数	5	6	3

观测集合 $V = \{红，白\}，M = 2$

状态集合 $Q = \{盒子1，盒子2，盒子3\}，N = 3$

$$\pi = (0.2, 0.4, 0.4)^T$$

$$A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

HMM建模

HMM系统特性由三个特征矢量或矩阵$\lambda = (A, B, \pi)$完全决定，其中A表示状态转移概率矩阵，B表示观测状态概率矩阵，π表示隐状态初始概率。

盒子	1	2	3
红球数	5	4	7
白球数	5	6	3

观测集合 $V = \{红，白\}$，$M = 2$

状态集合 $Q = \{盒子1，盒子2，盒子3\}$，$N = 3$

隐状态初始概率：$\pi = (0.2, 0.4, 0.4)^T$

状态转移概率矩阵：$A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$

观测状态概率矩阵：$B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$

HMM建模 – 三个基本问题

1. 概率计算问题（似然Likelihood ）

给定观测序列$O=\{O_1,O_2,\dots O_T\}$和模型$\lambda=(A,B,\pi)$， 如何高效计算产生观测序列的概率$P(O|\lambda)$?

- 前向后向算法

比如，观测序列O为（红，白，白），那么产生这一观测序列的概率是多少?

2. 学习问题（训练Learning ）

给出观测序列$\mathbf{O}=\{O_1,O_2,\dots O_T\}$，如何调整模型参数$\boldsymbol{\lambda}=(\boldsymbol{\pi},\mathbf{A},B)$，使得$\mathbf{P}(\mathbf{O}|\boldsymbol{\lambda})$最大?

- Baum-Welch参数估计算法

比如，估计A,B矩阵参数，使得重复三次抽球得到球的颜色的观测序列为（红，白，白）的概率最大。

3. 预测问题（解码Decoding ）

给定观测序列$O=\{O_1,O_2,\dots O_T\}$和模型$\lambda=(\pi,A,B)$， 如何选择最佳状态序列?

- Viterbi算法

比如，观测序列O为（红，白，白），那么最有可能从哪三个箱子抽出来的。

GMM-HMM语音识别系统流程（孤立词）

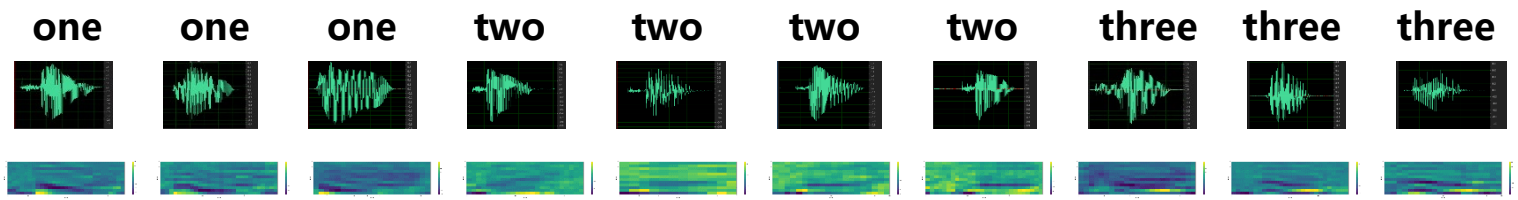
训练

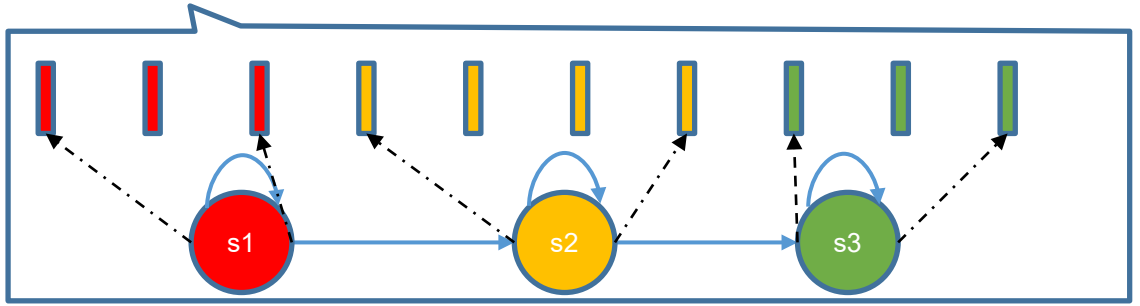
数据准备
音素，词典，
训练音频/文本

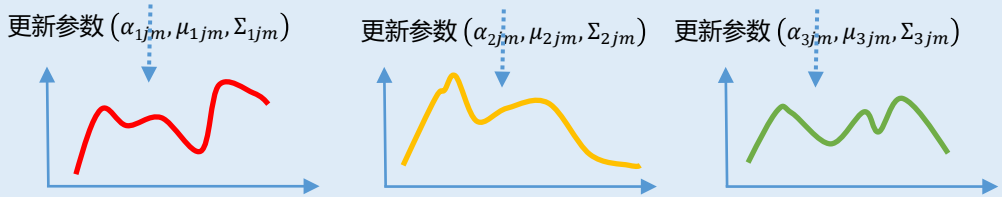
特征提取
MFCC

HMM状态
序列建模

GMM模型
概率密度建模



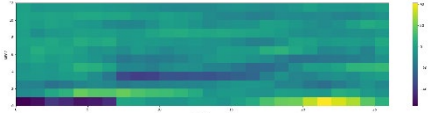




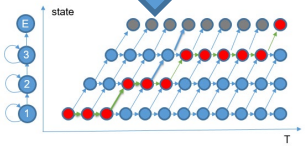
测试（解码）

未知
wav

提取
特征



Viterbi
解码图



解码
结果

two

HMM模型训练和解码流程

训练

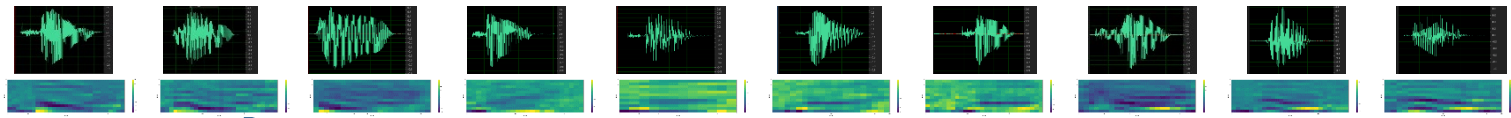
数据准备
抄本及对应的wav

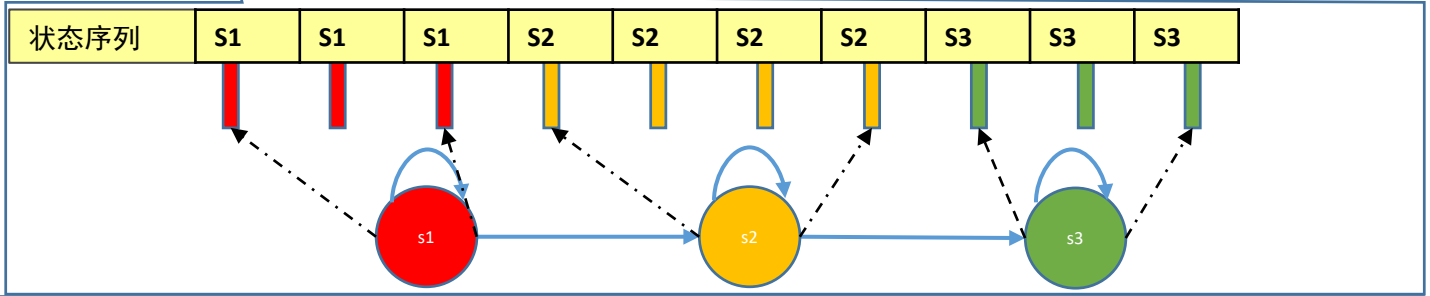
特征提取
激励参数/谱参数

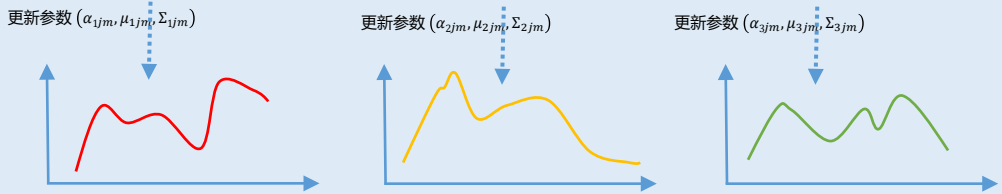
HMM状态
序列建模

GMM模型
概率密度建模

```
1 3500000 XX^XX-SIL+k=ao@X_X/A:X_X_X+X/B:X-X-X-X@X-X&X-X#X-X|X/C:4+X+2_X/D:X-X/E:X_X@X+X&X+X#X+X/F:2=1/G:X_X/H:X=X^X=X|X/I:4=3/J:16+10-4$
2 3500000 4431312 XX^SIL-k+ao=z@1_2/A:X_X_X+X/B:4-X=2-v@1-1&1-4#6-0|ao/C:3+uo+2_f/D:X-X/E:2_1@1+3&X+X#X+X/F:2=1/G:X_X/H:4=3^1=4|X/I:6=4/J:16+10-4$
3 4431312 5680483 SIL^k-ao+z=uo@2_1/A:X_X_X+X/B:4-X=2-v@1-1&1-4#0-2|ao/C:3+uo+2_f/D:X-X/E:2_1@1+3&X+X#X+X/F:2=1/G:X_X/H:4=3^1=4|X/I:6=4/J:16+10-4$
4 5680483 6121631 k^ao-z+uo=x@1_2/A:4_ao_2+v/B:3-X=2-f@1-1&2-3#2-0|uo/C:4+ia+2_v/D:2-1/E:2_1@2+2&X+X#X+X/F:4=2/G:X_X/H:4=3^1=4|X/I:6=4/J:16+10-4$
5 6121631 7306675 ao^z-uo+x=ia@2_1/A:4_ao_2+v/B:3-X=2-f@1-1&2-3#0-2|uo/C:4+ia+2_v/D:2-1/E:2_1@2+2&X+X#X+X/F:4=2/G:X_X/H:4=3^1=4|X/I:6=4/J:16+10-4$
```







预测（解码）

2. 基于隐马尔可夫(HMM)的统计参数语音合成

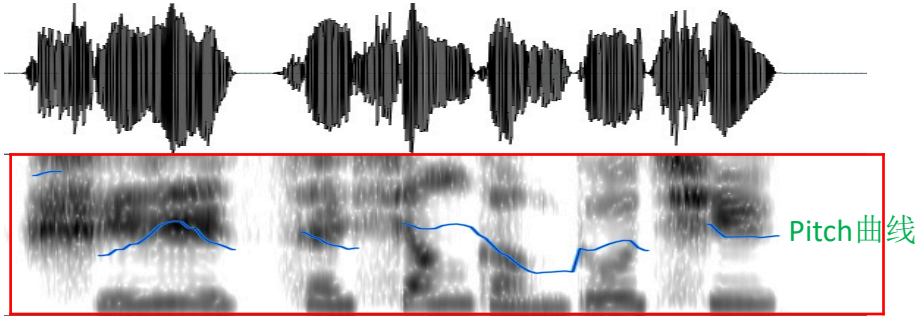
音频语音与语言处理研究组 SLP Audio, Speech and Language Processing Group NPU

- 2.1 统计参数语音合成框架
- 2.2 隐马尔可夫模型(HMM)
- **2.3 多空间概率分布MSD-HMM**
- 2.4 基于HMM的参数语音合成

HMM建模之F0：MSD-HMM

MSD-HMM：多空间概率分布HMM

multi-space probability distribution HMM



原因：

Pitch不连续，采用单流建模区分不出来清浊音，声音会很奇怪。

捕鱼问题：

池塘里有红鱼、蓝鱼、乌龟和垃圾，

- 捕捞上**红鱼**和**蓝鱼**的时候，我们对其**长和宽**感兴趣；
- 捕捞上**乌龟**的时候，对其**直径**感兴趣；
- 捕捞上**其他垃圾**的时候，则直接扔掉。

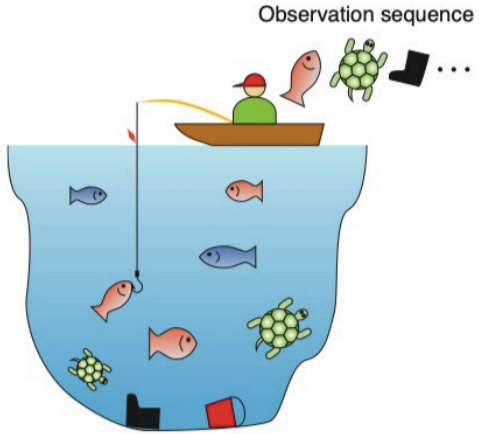


Figure 2 from An Introduction to HMM-Based Speech Synthesis

子空间space

样本空间Ω有G个子空间组成 $\Omega = \bigcup_{g=1}^{G} \Omega_g$ ，

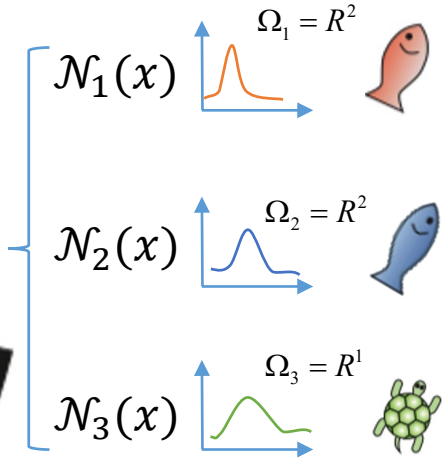
其中，Ω_g是一个n_g维的实空间R^{n_g}， g为空间索引值。

捕鱼问题包含了四个子空间

- Ω_1: 和红鱼相关的2 维空间，长度和宽度
- Ω_2: 和蓝鱼相关的2 维空间，长度和宽度
- Ω_3: 和乌龟相关的1 维空间，直径
- Ω_4: 和垃圾相关的0 维空间

$n_g > 0$，空间对应**分布**$\mathcal{N}_g(x)$，$x \in R^{n_g}$

$n_g = 0$，样本没有可衡量的维度



- 捞到红鱼时的观察o = ({1}, x)
- 假设晚上捞到鱼, 看不清楚颜色，那观察o是什么呢？ o = ({1, 2}, x)

空间权重

对每个空间Ω_g，都有一个相应的空间权重w_g。

样本

在此样本空间的样本o可以用一个观测矢量x以及对应的空间索引值X来表示，即

$$o = (X, x)$$

概率密度函数

其输出概率计算如下：

$$b(o) = \sum_{g \in X} w_g \mathcal{N}_g(x)$$

注意，当$n_g = 0$时，$\mathcal{N}_g(x) \equiv 1$。

也写成：$X = S(o)$，$x = V(o)$。

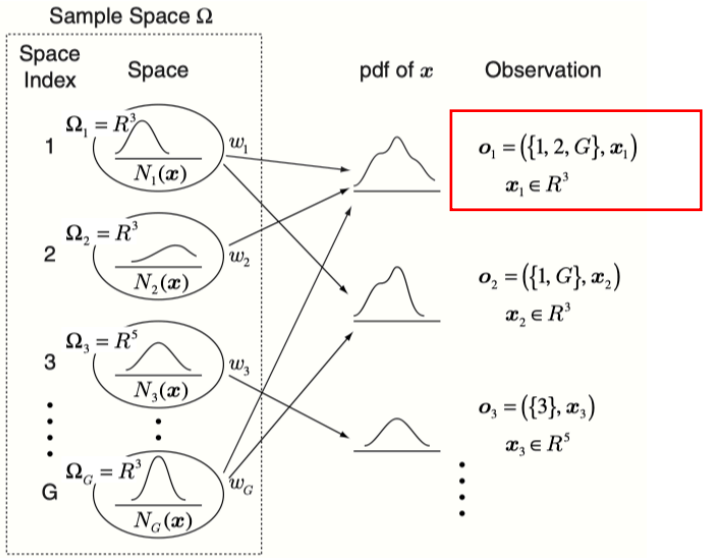


Figure from An Introduction to HMM-Based Speech Synthesis

对于o_1的 概率密度函数表示为：

$$b(o_1) = w_1 \mathcal{N}_1(x) + w_2 \mathcal{N}_2(x) + w_G \mathcal{N}_G(x)$$

MSD-HMM

MSD N-state HMM

- 初始化概率：

$$\pi = \{\pi_j\}_{j=1}^{N}$$

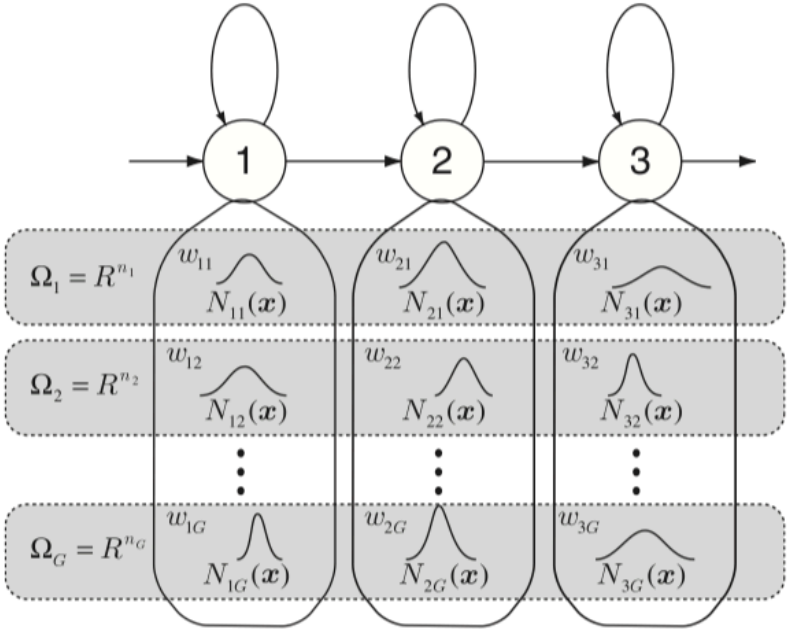
- 转移概率：

$$A = \{a_{ij}\}_{i,j=1}^{N}$$

- 输出概率：

$$B = \{b_i(\cdot)\}_{i=1}^{N}, \text{ where}$$

$$b_i(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_{ig} \mathcal{N}_{ig}(V(\boldsymbol{o})).$$



$$b_1(\boldsymbol{o}) = w_{11}\mathcal{N}_1(x) + w_{12}\mathcal{N}_2(x) + \cdots + w_{1G}\mathcal{N}_G(x)$$

Ω_1: 和红鱼相关的2 维空间，长度和宽度

Ω_2: 和蓝鱼相关的2 维空间，长度和宽度

Ω_4: 和垃圾相关的0 维空间

Figure from An Introduction to HMM-Based Speech Synthesis

MSD-HMM建模 声学特征

Pitch MSD-HMM 建模

G 取2，为两个子空间

- 空间1是对浊音建模，浊音是连续的实值。
- 空间2是对清音建模，轻音都是0。

$$S(o_t)=\begin{cases}\{1,2,\ldots,G-1\},(\text{浊音，}\ voiced)\\ \{G\},(\text{清音，}\ unvoiced)\end{cases}$$

	F_0 ave (Hz)	F_0 min (Hz)	F_0 max (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

谱参数和基频参数同时建模

- 第一个参数流为谱参数，包括静态参数、一阶和二阶差分参数；
  采用**连续概率分布**进行建模。
- 第二到第四个参数流分别为基频参数的静态值、一阶差分、二阶差分参数；
  采用**多空间概率分布**建模。

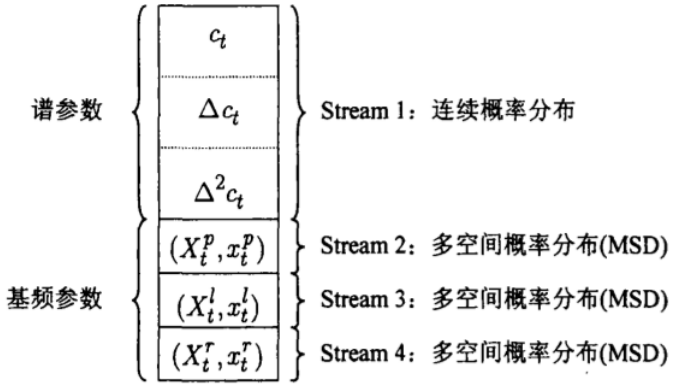


Figure from 基于隐马尔科夫模型的语音合成技术研究

几个概念：多维 混合高斯 多流

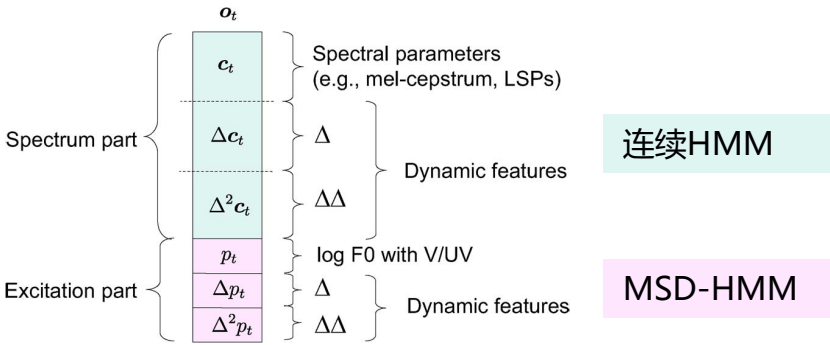
1. 多维高斯

在单音素HMM模型中，每个状态对应大量的**多维**的观测值

包括：静态特征：F0， 谱参数等

动态特征：delta F0， delta delta F0 等

- 需要用一个**多维**高斯模型对整个状态到观测进行建模。



2. 混合高斯（GMM）

- ASR中，需要考虑不同人不同环境的语音参数变化，采用混合高斯数目比较多，比如10～20。
- TTS中，由于基本是单一speaker和上下文相关的HMM建模，一般采用1个混合高斯分量。

- 单个高斯分布：

$$b_j(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

- 高斯混合模型：

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} \alpha_{jm} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$$

其中 $j = 1,2,\dots,N;\ m = 1,\dots,M$ 表示GMM分量编号

3. 多流（MSD）

- 其中一维（如F0）又分成多个子空间（清音+浊音）

$$b_1(\boldsymbol{o}) = w_{11}\mathcal{N}_1(x) + w_{12}\mathcal{N}_2(x) + \cdots + w_{1G}\mathcal{N}_G(x)$$

几个概念：多维 混合高斯 多流

音频语音与语言 SLP 处理研究组 Audio, Speech and Language Processing Group NPU

混合高斯分布GMM

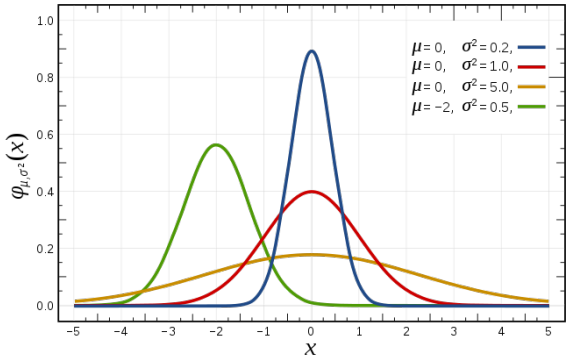
单高斯分布（多维） $\xrightarrow{D=1}$ 一维高斯分布

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}.$$

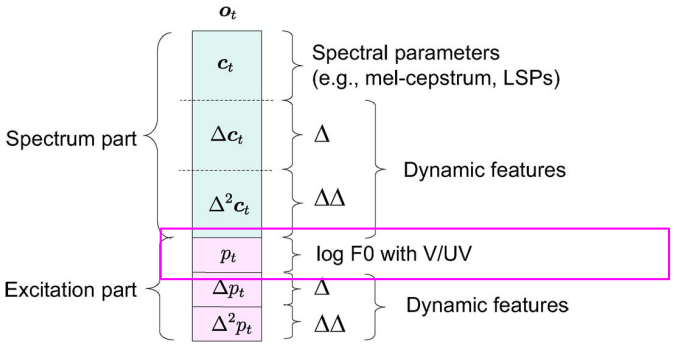
$$X \sim N(\mu,\sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



多空间概率分布MSD

- 激励参数+谱参数
- 多个参数，每个参数再分成多个**子空间（多流）**
- 子空间不一定用高斯建模



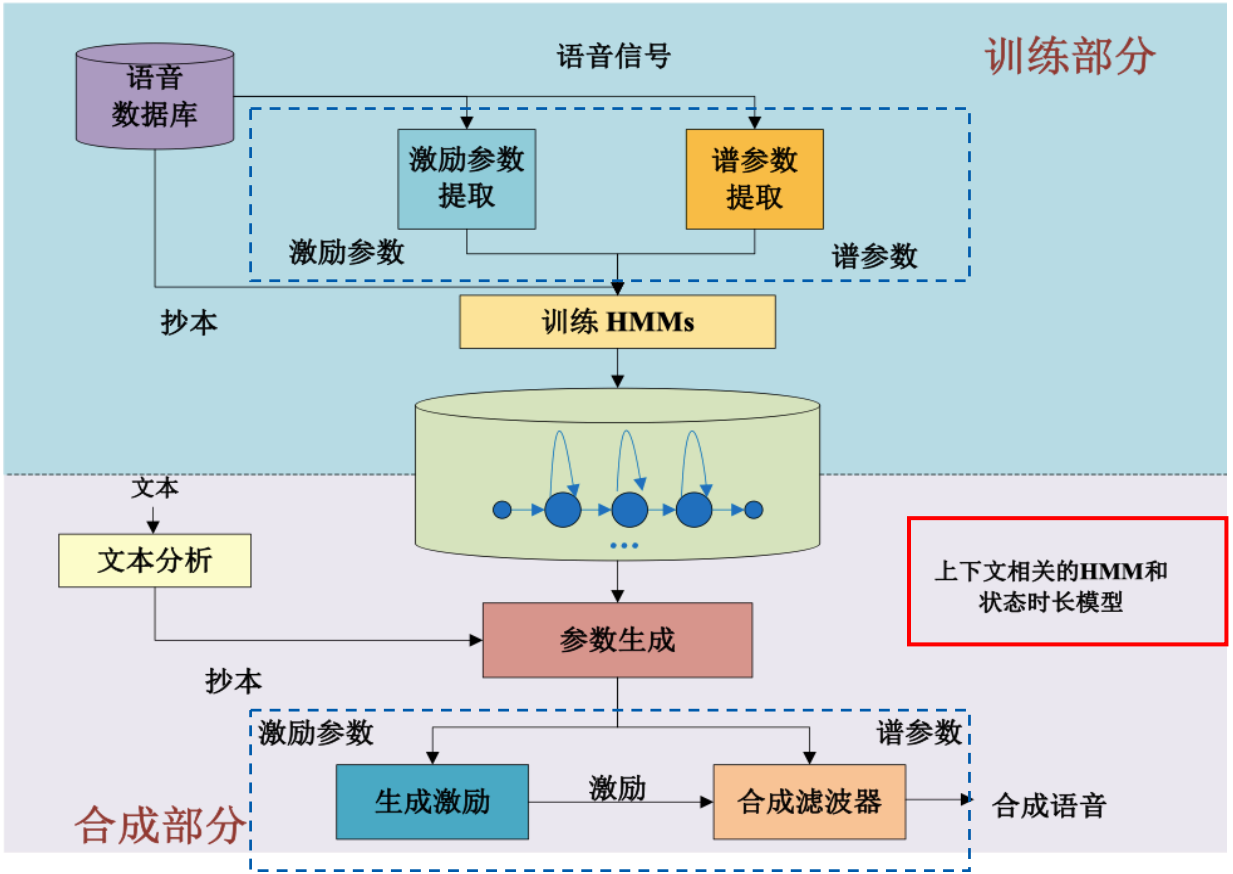
$$S(o_t) = \begin{cases}\{1,2,\ldots,G-1\}, (\text{浊音}, voiced)\\ \{G\}, (\text{清音}, unvoiced)\end{cases}$$

2. 基于隐马尔可夫(HMM)的统计参数语音合成

音频语音与语言处理研究组 SLP Audio, Speech and Language Processing Group NPU

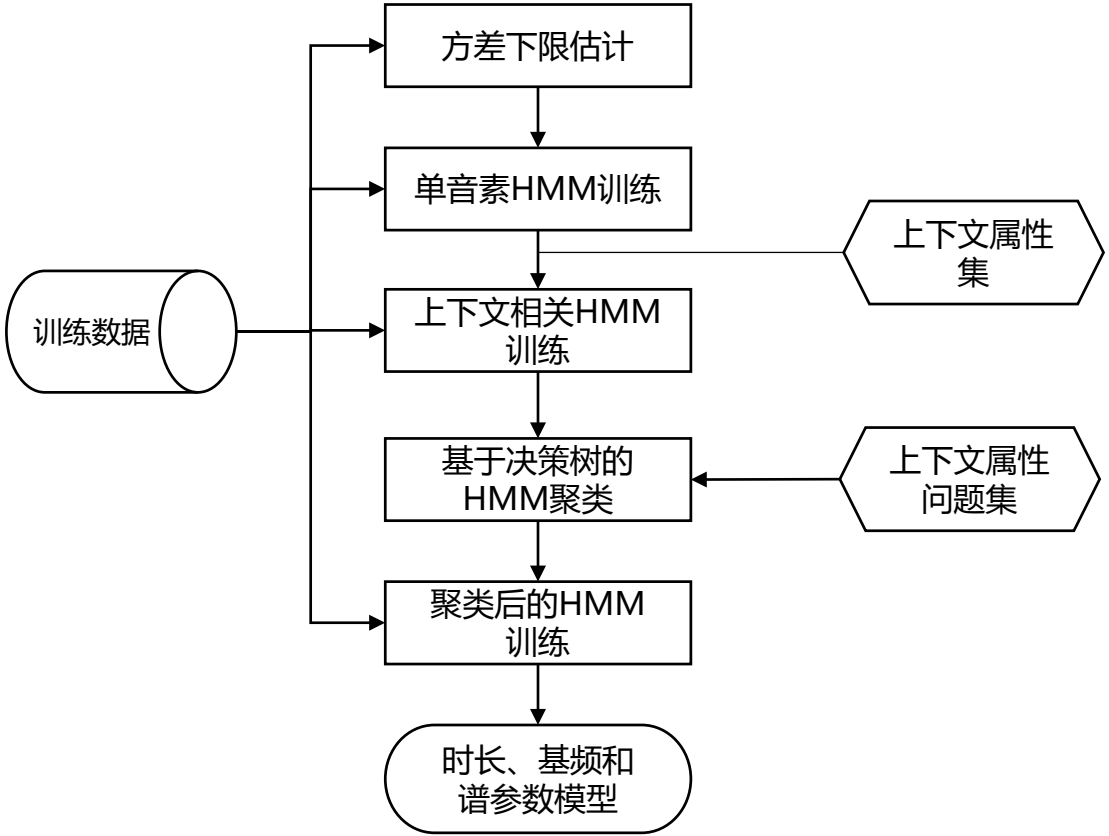
- 2.1 统计参数语音合成框架
- 2.2 隐马尔可夫模型(HMM)
- 2.3 多空间概率分布MSD-HMM
- **2.4 基于HMM的参数语音合成**

基于HMM的参数语音合成整体流程



HMM建模之训练

训练流程



HMM建模之决策树

为什么要用决策树?

- 输入特征过于复杂
- 训练数据不够多
- 如果不用决策树，导致模型过拟合

决策树例子如右图

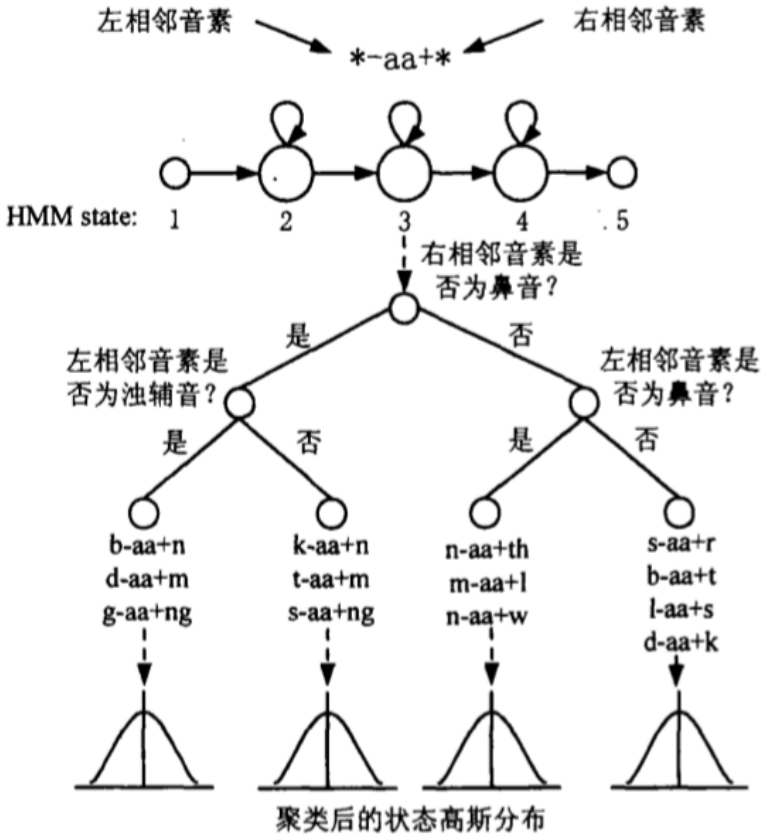
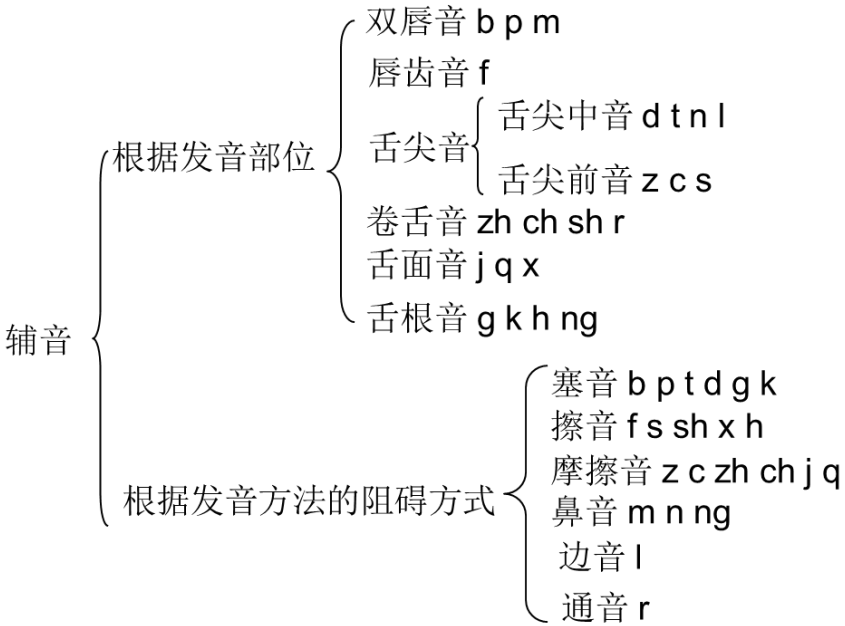


Figure from 基于隐马尔科夫模型的语音合成技术研究

HMM建模之决策树

设计

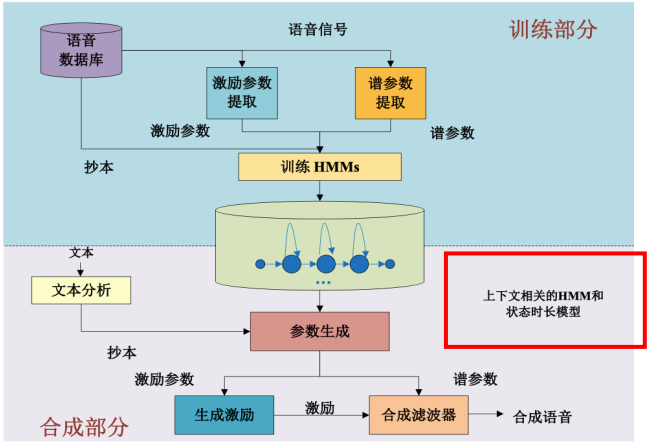


QS "LL_Initial_Place_ShuangChun" {b^*,p^*,m^*} 左边的左边的声母是双唇音
QS "LL_Initial_Place_ChunChi" {f^*} ……是唇齿音
QS "LL_Initial_Place_SheJian" {d^*,t^*,n^*,l^*,z^*,c^*,s^*} ……是舌尖音
QS "LL_Initial_Place_SheJianZhong" {d^*,t^*,n^*,l^*} ……是舌尖中音
QS "LL_Initial_Place_SheJianQian" {z^*,c^*,s^*} ……是舌尖前音
QS "LL_Initial_Place_JuanShe" {zh^*,ch^*,sh^*,r^*} ……是卷舌音
QS "LL_Initial_Place_SheMian" {j^*,q^*,x^*} ……是舌面音
QS "LL_Initial_Place_SheGen" {g^*,k^*,h^*} ……是舌根音
QS "LL_Initial_Block_Blocked" {b^*,p^*,t^*,d^*,g^*,k^*} ……是塞音
QS "LL_Initial_Block_Fricate" {f^*,s^*,sh^*,x^*,h^*} ……是擦音
QS "LL_Initial_Block_BloFri" {z^*,c^*,zh^*,ch^*,j^*,q^*} ……是摩擦音
QS "LL_Initial_Block_Nasal" {m^*,n^*} ……是鼻音
QS "LL_Initial_Block_Bian" {l^*} ……是边音
QS “LL_Initial_Block_Tong” {r^*} ……是通音

时长模型(duration model)

时长建模

- 采用多维单高斯模型
- 高斯维度是HMM的状态个数



时长展开

谱参数

上下文相关时长模型

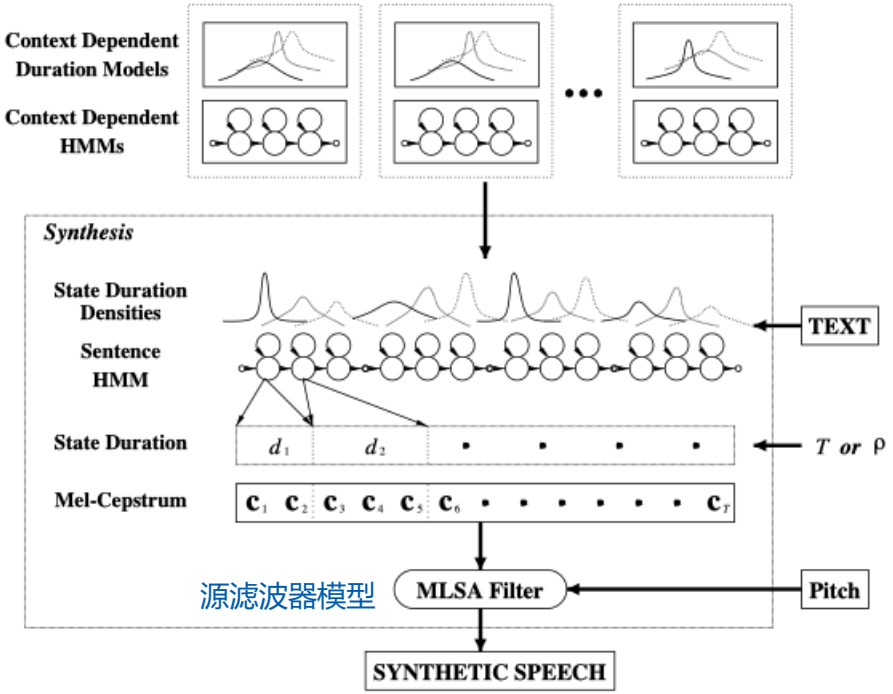


Figure from DURATION MODELING FOR HMM-BASED SPEECH SYNTHESIS

时长模型(duration model)

在句子中，通常的做法是将每个词对应的音素模型从左到右连接起来构造HMM网络。

右图是a, i两个音素构造的HMM网络，在合成的时候需要决定：

- 每个状态对应的观测
- 每个音素对应多少个状态（duration模型）

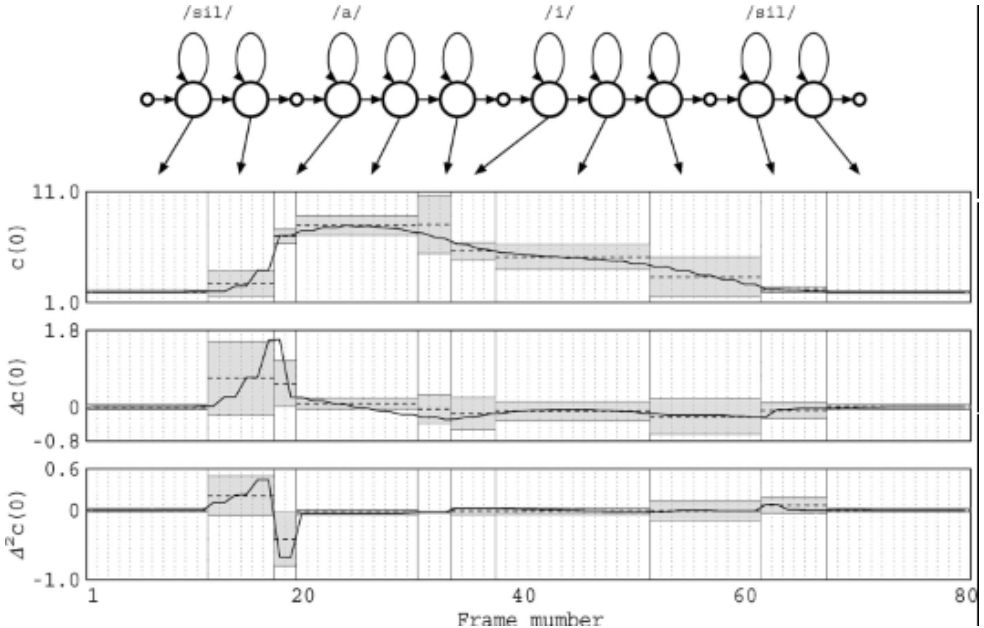
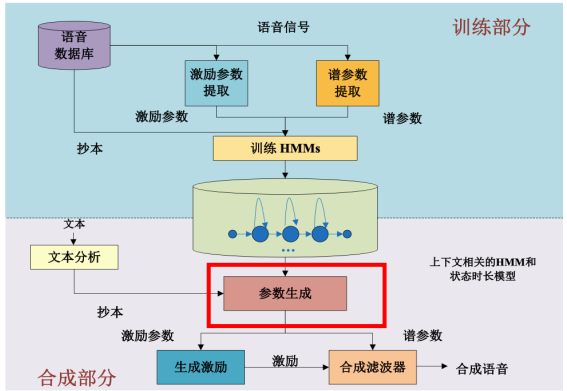


Figure from http://vsooda.github.io/2017/02/23/tts-tutorial/

HMM建模之参数生成算法



训练好的模型参数　　语速相关

$$P[O \mid \lambda, T] = \sum_{Q} P(O \mid Q, \lambda, T) P(Q \mid \lambda, T)$$

$$\cong \max_{Q} P(O \mid Q, \lambda, T) P(Q \mid \lambda, T)$$

谱参数，基频　　状态模型，时长

$$Q_{\max} = \arg\max_{Q} P(Q \mid \lambda, T)$$

$$O_{\max} = \arg\max_{O} P(O \mid Q_{\max}, \lambda, T)$$

对于时长模型 Q_{max}

要调整的时长

$$P(Q \mid \lambda, T) = \prod_{i=1}^{K} p_i(d_i)$$

均值　　时长缩放因子（调速）

$$d_k = m_k + \rho \cdot \sigma_k^2, \qquad 1 \le k \le K,$$

$$\rho = \left(T - \sum_{k=1}^{K} m_k \right) \Big/ \sum_{k=1}^{K} \sigma_k^2,$$

高斯分布对时长建模
K为状态时长个数, ρ为总的时长缩放音子

对于声学参数模型 O_{max}

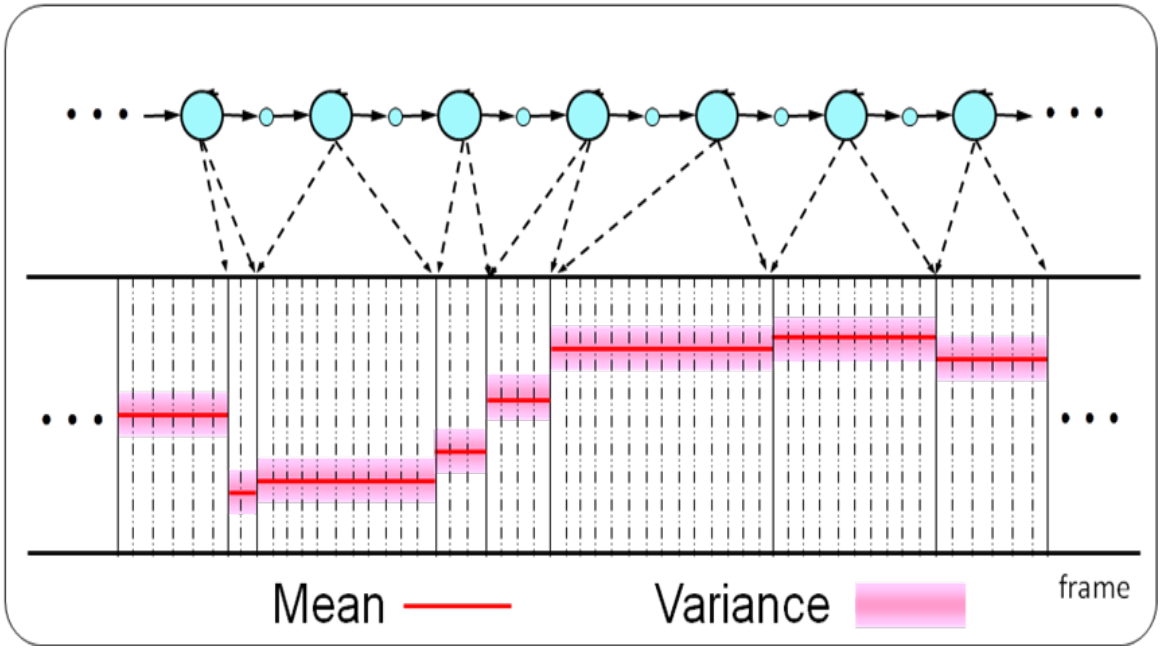
$$P(O \mid Q_{\max}, \lambda, T) = \prod_{t=1}^{T} b_{q_t}(o_t)$$

$$o_t = \arg\max_{o} b_{q_t}(o), \qquad t = 1, 2, \ldots, T$$

q为状态
只考虑静态特征，最大值为高斯分布的均值

HMM建模之动态参数生成算法

只有静态特征，频谱不连续



HMM建模之动态参数生成算法

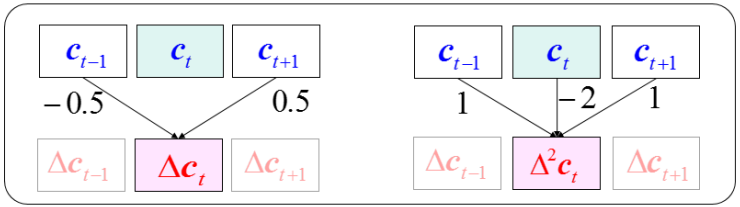
动态特征：如右图

动态参数生成算法

- 参考(An Introduction to HMM-Based Speech Synthesis)
- 算法实现(SPRK-mlpg)

$$\Delta \boldsymbol{c}_t = \frac{\partial \boldsymbol{c}_t}{\partial t} \approx 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1})$$

$$\Delta^2 \boldsymbol{c}_t = \frac{\partial^2 \boldsymbol{c}_t}{\partial t^2} \approx \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1}$$



NAME

mlpg – obtains parameter sequence from PDF sequence[23]

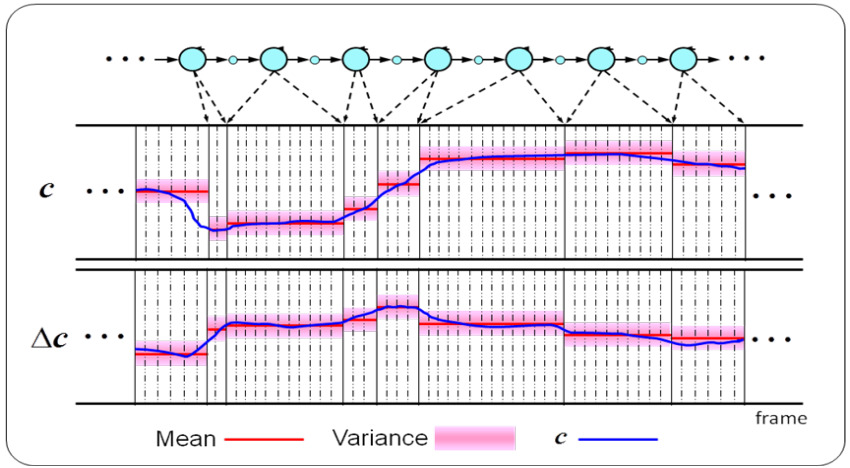
SYNOPSIS

mlpg [**–l** *L*] [**–m** *M*] [**–d** (*fn* | d_0 [d_1 ...])] [**–r** N_R W_1 [W_2]] [**–i** *I*] [**–s** *S*] [***infile***]

DESCRIPTION

mlpg calculates the maximum likelihood parameters from the means and diagonal covariances of Gaussian distributions from *infile* (or standard input), and sends the result to standard output. The input format is

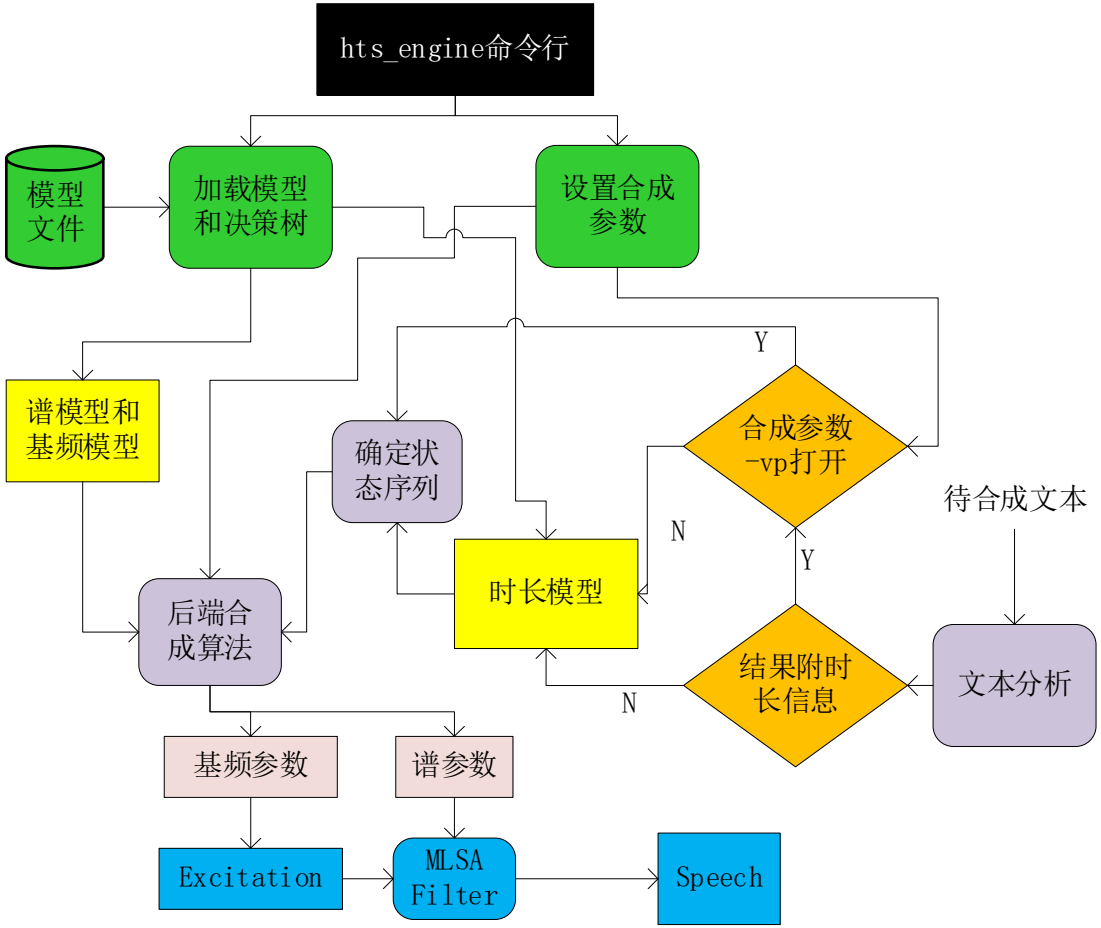
$$\ldots, \mu_t(0), \ldots, \mu_t(M), \mu_t^{(1)}(0), \ldots, \mu_t^{(1)}(M), \ldots, \mu_t^{(N)}(M),$$
$$\sigma_t^2(0), \ldots, \sigma_t^2(M), \sigma^{(1)2}{}_t(0), \ldots, \sigma^{(1)2}{}_t(M), \ldots, \sigma^{(N)2}{}_t(M), \ldots$$



HMM语音合成整体流程

Hts_engine

- 基于HMM的语音合成工具包
- http://hts.sp.nitech.ac.jp/



目录

语音合成方法：统计参数合成（第一讲）

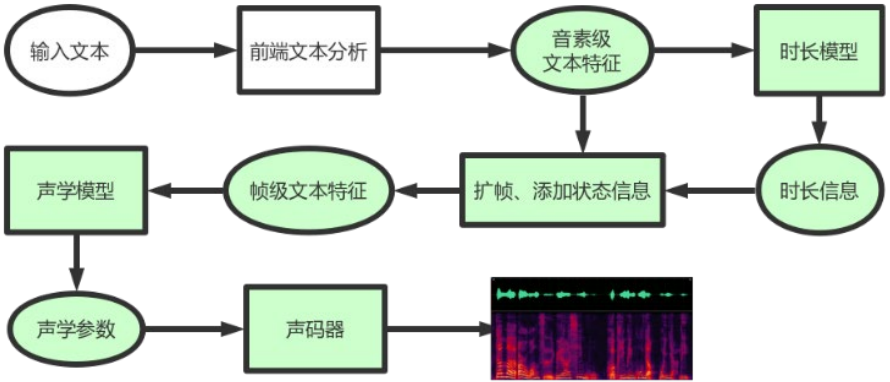
文本→语音: (非常)不等长序列映射

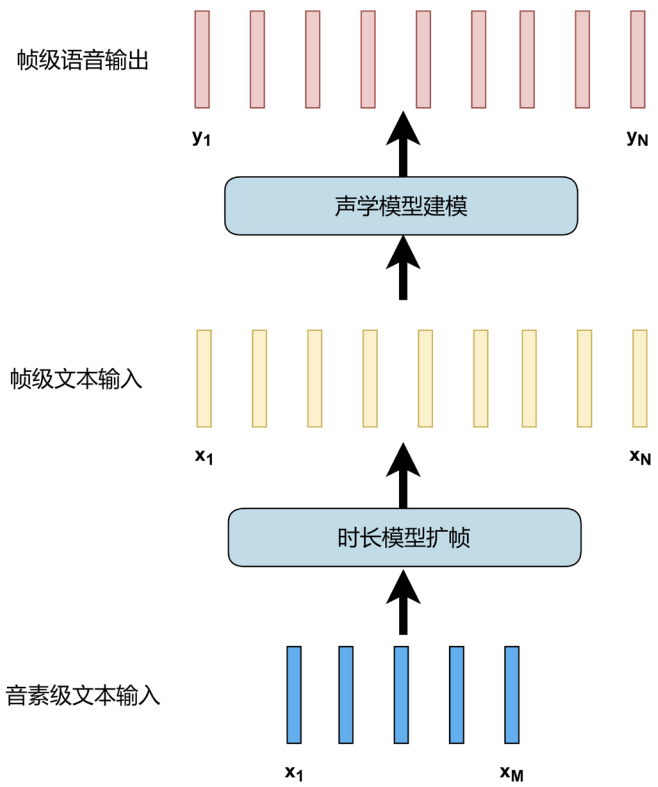
帧级建模

- 时长模型：音素序列→帧级文本特征
- 声学模型：帧级文本特征→帧级语音输出

训练数据

- 利用语音识别强制对齐，得到音素帧级对应关系





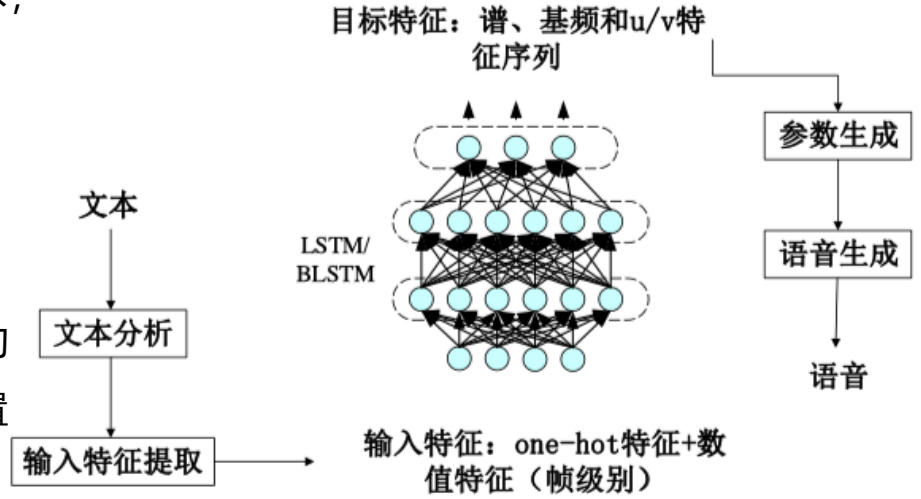
基于NN的参数语音合成

时长模型

- 输入：经过文本分析从HMM系统中的抄本，转换出的one-hot 特征 + 数值特征
- 输出：音素时长+其状态时长

声学模型

- 输入：时长模型的输入根据时长转换得到的frame级别信息 + 状态以及帧数相关的位置信息
- 输出：谱参数、基频、清浊音(u/v)
- 输出再经过声码器，得到语音



基于NN的参数语音合成

输入特征表示

- 将数字表示的特征进行保留，如“当前音节所包含的音素数目 =5”，则在神经网络的输入的对应维度写上5；
- 将类别的特征使用one hot 表示法，即用和类别数目一样维度的向量表示该特征，比如当前音素为b，b 在音素列表中是第2个， 总共音素66个，我们就在第2维填上1，其他65维填上0；
- 增加状态信息，如“当前帧处于第5 个状态”，总共用7 状态建 模，则将状态信息表示为“0 0 0 0 1 0 0”。

对于时长模型

输入是音素对应的特征表示，输出是其时长+对应状态时长。一句话有多少音素就对应有多少输入。

对于参数生成模型

输入是**frame**级别信息，根据音素时长，展开为frame级别输入，并且增加状态信息。所以一句话有多少frame就有多少输入，和输出语音特征对应。

基于NN的参数语音合成

HMM合成的过平滑问题导致合成效果平淡，音质不好。

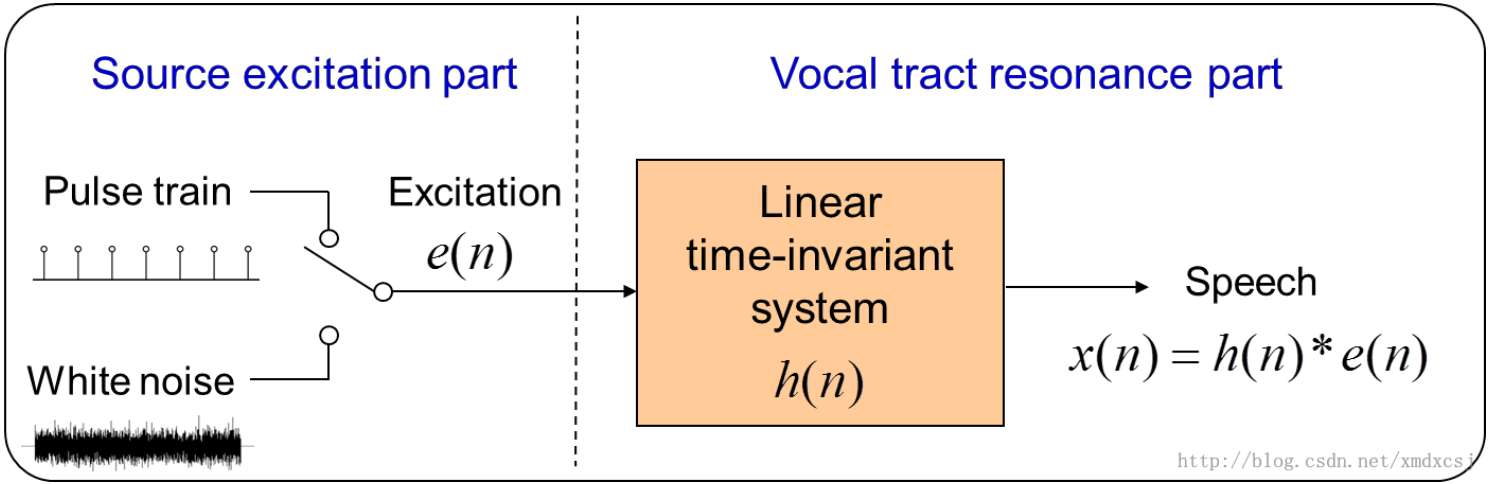
- HMM基于建模，假设各状态之间独立，状态参数分布式统计平均的结果；
- 决策树作为浅层线性模型无法建模复杂的映射关系，对输入特征和数据空间的线性分割会降低模型泛化能力；
- 特征的短时相关特性。

使用深度神经网络(LSTM-RNN)替代决策树直接对文本特征和声学特征映射关系进行建模。

- 基于帧建模替代HMM状态建模，可直接输出到声码器；
- 多层神经网络替代浅层的决策树。

目录

源-滤波器模型(Source-filter model)



源-滤波器模型认为：声音是由激励(excitation)和相应的滤波器(vocal tract)组成。

声道模型(vocal tract)

- 将声道看做一个谐振腔，“共振峰模型”

源-滤波器模型(Source-filter model)

声带激励分为两类，可以产生清音或者浊音。

- 浊音（voiced）

气流通过紧绷的声带，对声带进行冲击而产生振动，使声门处形成准周期性的脉冲串，激励信号简化为周期性的脉冲激励。

- 清音（unvoiced）

声带处于松弛状态，不发生振动，气流通过声门直接进入声道，激励信号简化为随机白噪声。

上面的二元激励模型将复杂的产生激励过程简单的划为两部分，大大简化了声门激励的特征，但是合成语音的自然度较低。

参数对应关系

- F0/pitch对应于激励部分的e(n)的周期脉冲序列和白噪声的叠加
- spectral envelope对应于声道谐振部分的h(n)
- aperiodicity对应于激励部分的e(n)中的非周期序列

工具

- HTS vocoder : http://hts.sp.nitech.ac.jp/
- Straight : https://github.com/HidekiKawahara/legacy_STRAIGHT
- World: https://github.com/mmorise/World

目录

单元拼接语音合成

目的：

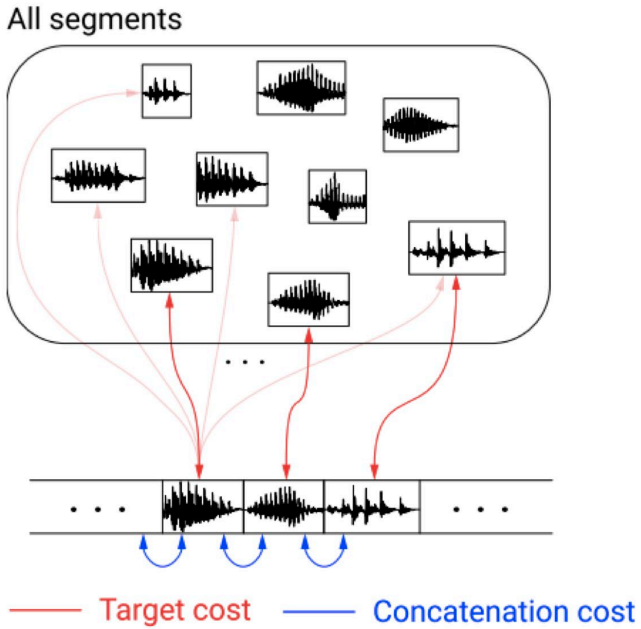
拼接生成一段很自然的声音

方法：

从录制好的音频数据库中选择拼接单元进行拼接

拼接方式：

- phoneme synthesis
- diphone synthesis
- syllable synthesis
- word synthesis
- ...



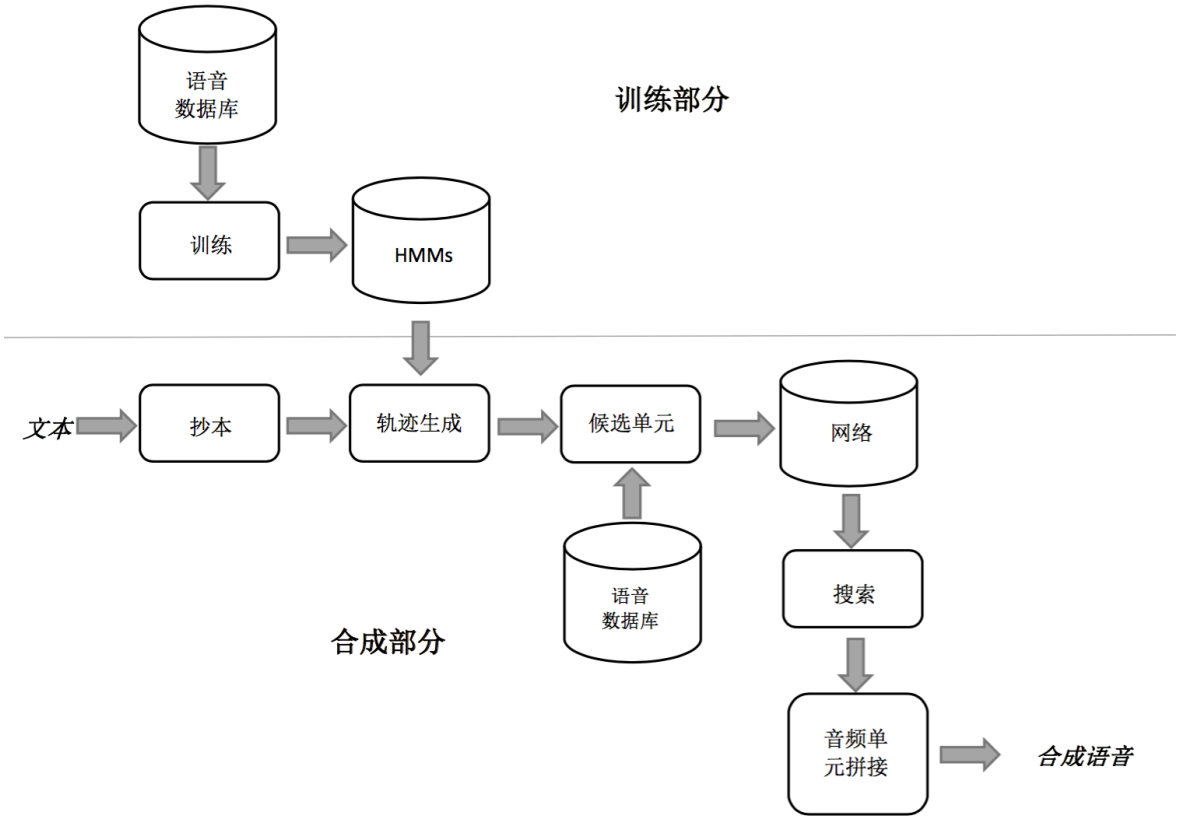
单元拼接语音合成

训练阶段

- 数据准备与预处理
- 模型训练
- 模型修正
- 音库建立

合成阶段

- 前端文本分析
- 参数轨迹预测
- 候选单元预选
- 搜索网络构建
- 最优路径搜索
- 波形拼接



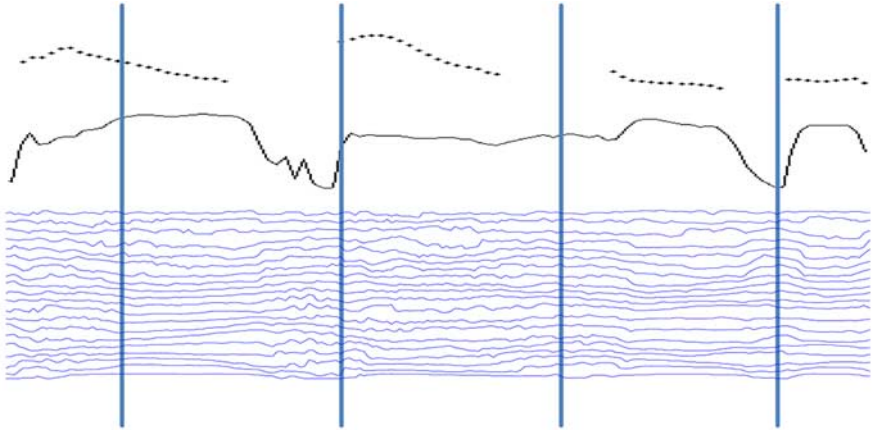
单元拼接语音合成（轨迹）

音频语音与语言处理研究组 SLP Audio, Speech and Language Processing Group NPU

示意图

Guiding parameter trajectories

轨迹



- **生成轨迹**

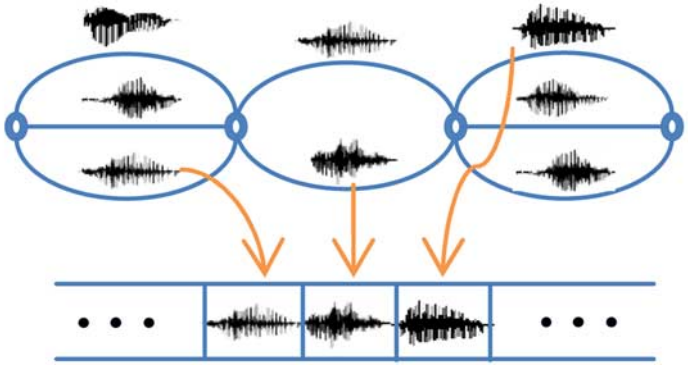
F0 基频

Gain 音量

LSP 谱参数

“Sausage” of waveform tiles

Waveform tile concatenation



- **划分网格**
- **不能硬拼**

单元拼接语音合成

音库准备与建立

- Key-value模式存储
- Key：音素+音调+个数 w_o_sh_3_123
- Value：音频采样点、谱参数、基频

HMM/LSTM 轨迹生成

- 得到每个音素对应的时长、谱参数、基频、增益

单元筛选及网络构建

- 依赖label的上下文，选择出相同的label（三音素、双音素或者单 因素）作为候选单元
- 依据下面公式，得到谱参数、基频、增益的距离（此距离也作为目标代价），按照预先设定好的阈值进行筛选
- 时长的筛选，5-10帧的差别
- 排序候选单元，选择最优的前N个单元
- 每个音素都会有很多候选单元，组成网络

$$d_{F0} = |log(F0_t) - log(F0_c)|$$

$$d_G = |log(G_t) - log(G_c)|$$

$$d_w = \sqrt{\frac{1}{I}\sum_{i=1}^{I} w_i(w_{t,i} - w_{c,i})^2}$$

$$w_i = \frac{1}{w_{t,i} - wt, i-1} + \frac{1}{w_{t,i+1} - w_{t,i}}$$

单元拼接语音合成

拼接代价

- normalized cross-correlation(NCC)，归一化互相关

$$r(d) = \frac{\sum_t [(x(t) - \mu_x) \cdot (y(t) - \mu_y)]}{\sqrt{\sum_t (x(t) - \mu_x)^2} \cdot \sqrt{\sum_t (y(t) - \mu_y)^2}}$$

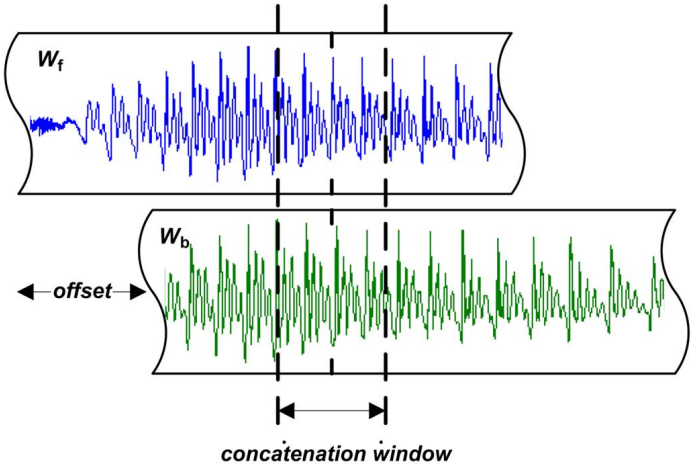
- 前后单元帧之间的距离

最优路径

- 维特比算法

单元拼接

- NCC得到最优拼接位置
- 三角窗，淡入淡出进行拼接



基于HMM/NN的语音合成、拼接语音合成对比

HMM语音合成	
优点	**缺点**
训练速度快	音质不好，有机械感
模型小	模型平均，没有准确预测参数
合成速度非常快	依赖源滤波器vocoder

NN语音合成	
优点	**缺点**
神经网络更好的拟合参数	仅仅预测语音参数的均值，没有多样性
能够拟合语音长依赖的特性	时长模型和声学模型分开建模
合成速度快	依然使用源滤波器声码器

拼接语音合成	
优点	**缺点**
声音自然流畅	需要大量的数据库
能够还原发音人的真实声音	消耗内存/磁盘 大
	模型多样性少

实践3：基于LSTM/GRU的声学与时长模型

尝试按照README.md中的步骤，完成模型构建、训练、测试

- 声码器
  - 测试使用world声码器进行copy synthesis （必做）
  - 测试使用griffinlim进行copy synthesis （选做）
- 声学、时长模型
  - 在现有代码基础上完成模型部分代码，参考readme中训练测试过程，使用给出的训练集训练模型，在给定测试集上进行测试合成。（必做）

Repo：

https://github.com/nwpuaslp/TTS_Course

深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

感谢聆听！

Thanks for Listening