语音合成
Speech Synthesis

清华大学深圳研究生院
吴志勇
zywu@sz.tsinghua.edu.cn

MEDIA JOINT Research Center

Overview

- **Speech Synthesis**

- **Text-to-Speech Synthesis: TTS**
  - Architecture of TTS

- **TTS in Detail**
  - Text Analysis
  - Prosodic Analysis
  - Unit Selection
  - Prosodic Modification

Speech Synthesis

- **Speech synthesis**
  - Speech synthesis is the artificial production of human speech.
- **语音合成: 计算机话语输出**
  - 让计算机象人那样讲话。
  - 语音合成的研究目标是：
    - 可懂、清晰、自然、具有表现力。
  - 人们用语言进行交流时，用声音来表达**事实**，也表达**意向**、**情感**。计算机也应该像人那样讲话。

某人问你：你愿意和我一起去看电影吗？
你的回答可能是：
“是的，我很高兴和你一起去看电影。”（**肯定，高兴**）
“抱歉，我不能和你一起去看电影，因为我要去开会。”（**无可奈何**）
“不去，还是你自己去看吧。”（**否定**）

The Voice of Stephen Hawking

- The voice of Stephen Hawking

SPEAK WRITING
1. Speak sentence shown below.
2. Go back one sentence.
3. Skip to next sentence.
4. Go to beginning of speech.
5. Go to TALKING Work Area
6. Return to Main Menu.
Yes
No
Maybe
I don't know
Thank you
d a great deal of interest but many p
e found it difficult to understand. I
efore decided to try to write a new
on that would be easier to follow

A Brief History of Time

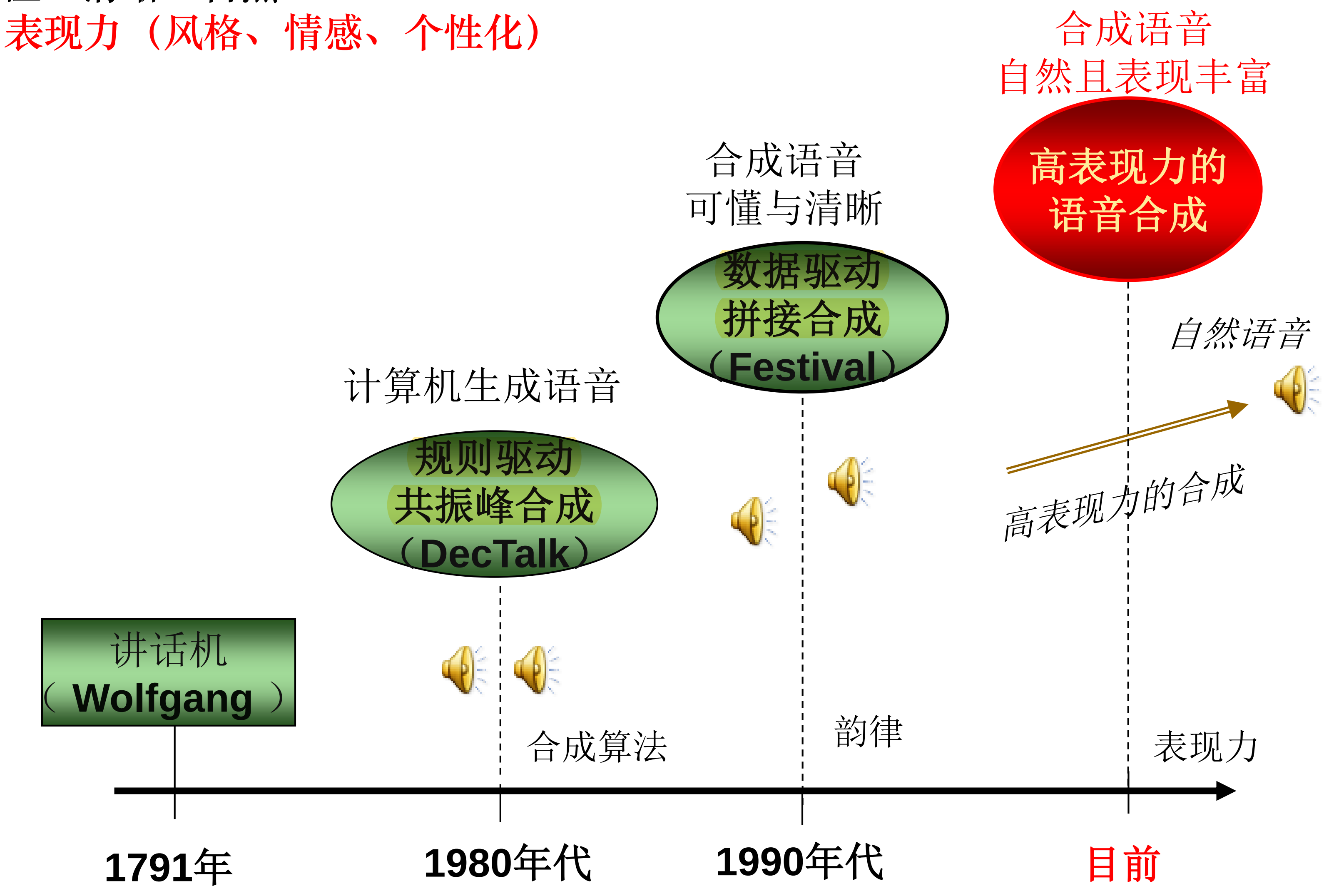
Research Objective

- **研究目标：让计算机象人那样讲话**
  - 确保可懂度（1982年） 🔊
  - 提高清晰度（1984年） 🔊
  - 改善自然度（1992年—） 🔊 🔊
  - 具有情感、表现力（？） 🔊

A Brief History

可懂→清晰→自然

➢表现力（风格、情感、个性化）



Speech Production: 语音产生

- **Functions of the Articulators: 各发音器官的功能**
  - **肺** (lungs): 发音气流源头
  - **声带** (vocal cords, vocal folds, larynx): 受气流影响相互靠近收紧，发生振动，产生浊音 (voiced); 或者声带松弛使声门 (glottis) 开放，产生清音 (voiceless / unvoiced)
  - **软腭** (soft palate, velum): 具有阀门功能，打开时允许气流进入鼻腔 (nasal cavity)，关闭时禁止气流进入鼻腔
  - **硬腭** (hard palate): 口腔 (oral cavity) 顶部较长的硬表面。舌顶住硬腭时，发辅音 (consonant)
  - **舌** (tongue): 灵活的发音器官，远离硬腭发元音 (vowel) ; 靠近或接触硬腭或其他硬表面发辅音
  - **牙齿** (teeth): 发某些辅音时，用来顶住舌
  - **唇** (lips): 变圆或扁影响发元音的质量，或者完全紧闭，阻止气流从口腔发出

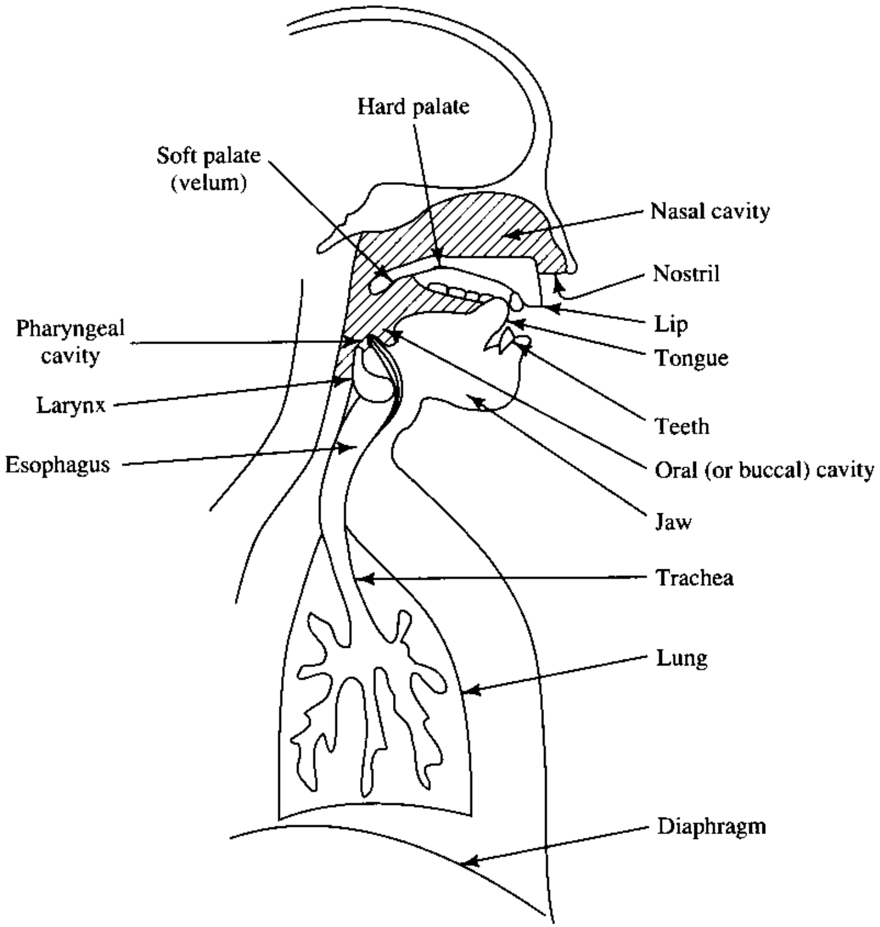


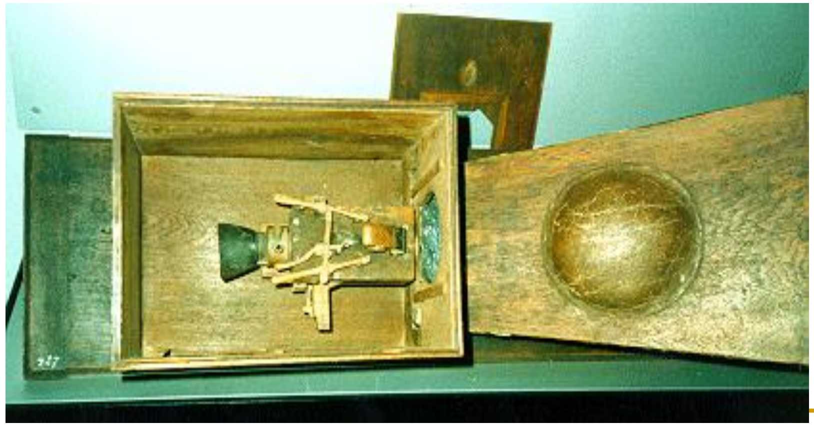
Diagram of the articulators (speech organs)

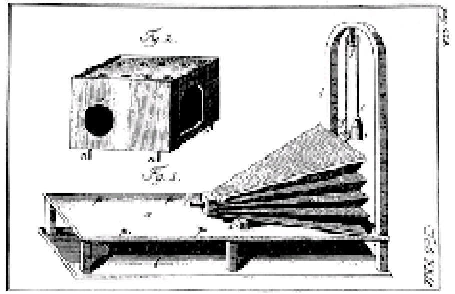
(from: http://imp.lss.wisc.edu/~jrvalent/AIS/Grammar/Phonology/Phonol002a.html)

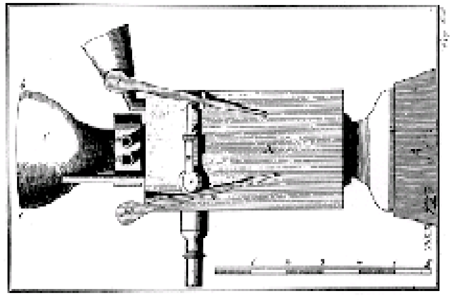
Predecessor of Speech Synthesizer

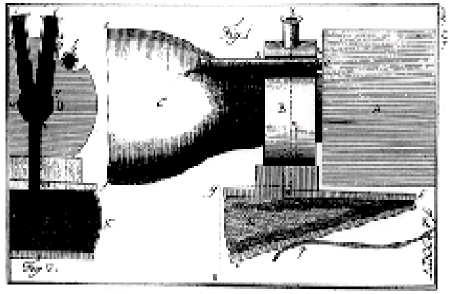
- **Speak Machine, 1769-1790 – Wolfgang von Kemplelen**
  - 第一台机械发声的语音器
  - 1791年，匈牙利科学家Wolfgang von Kempelen利用一系列精巧的风箱、弹簧片、风笛与共鸣箱制造出一个可以发出简单词汇的机器，取名为“讲话机”。

http://www.ling.su.se/staff/hartmut/kemplne.htm



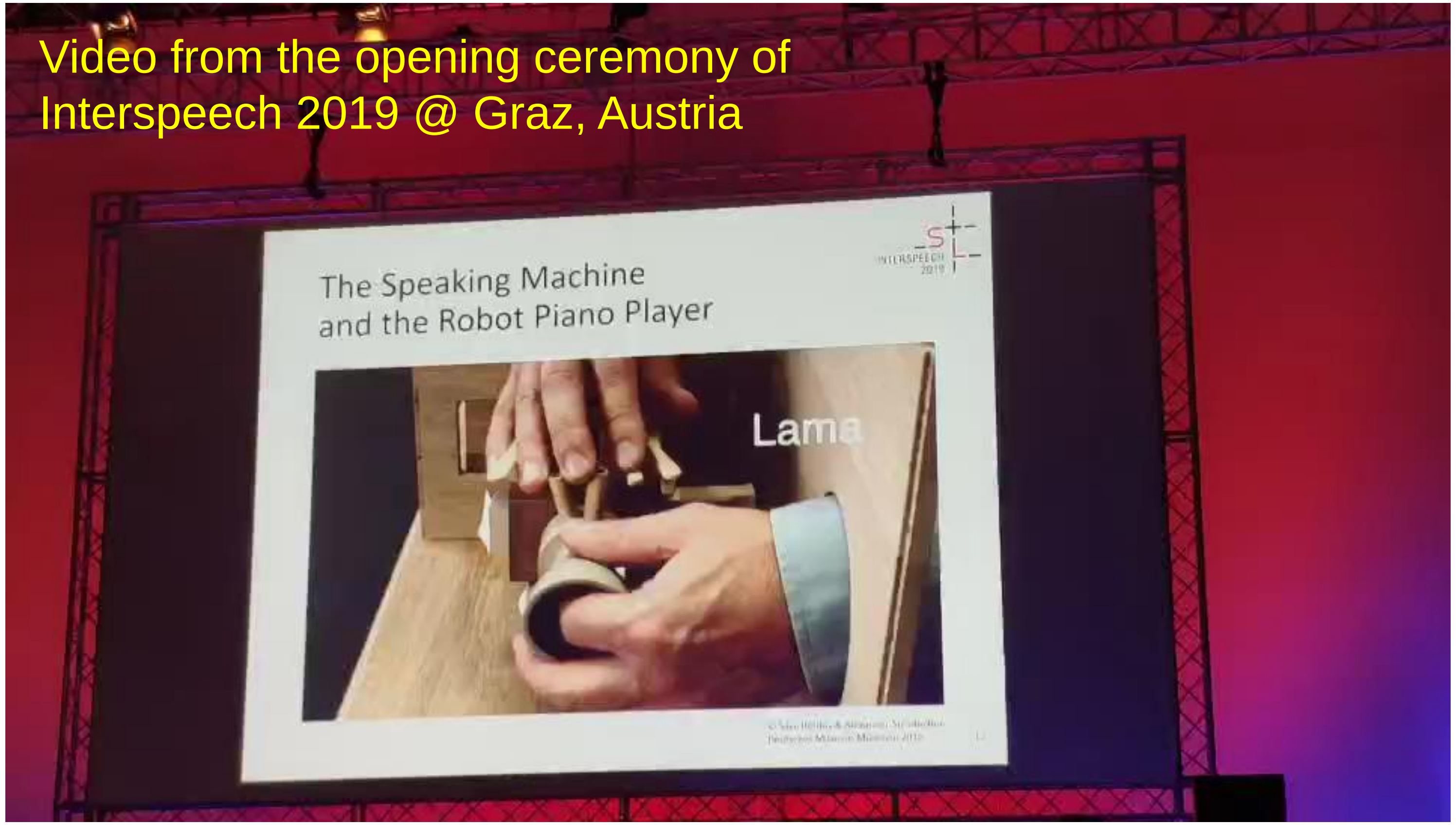


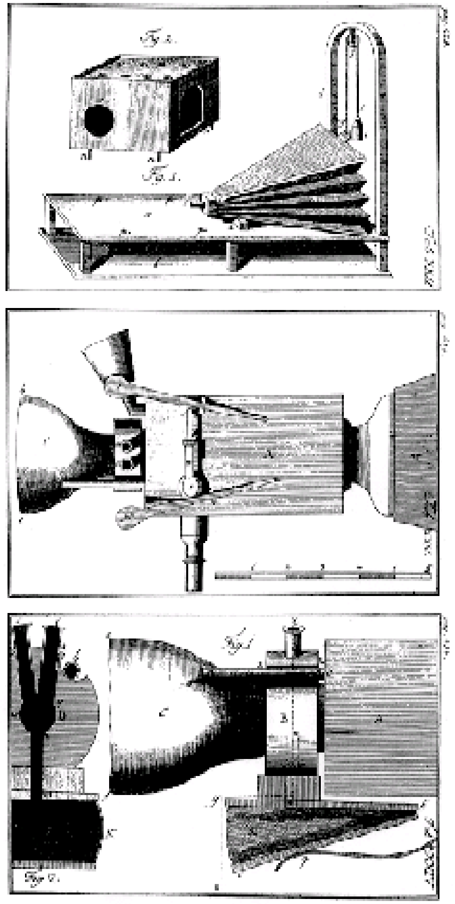




Predecessor of Speech Synthesizer

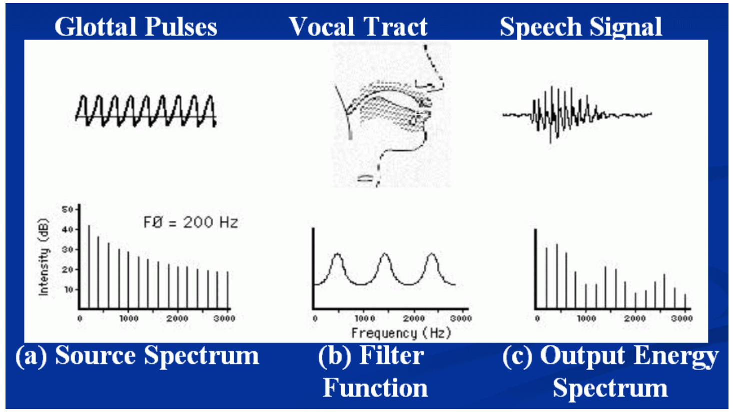
- **Speak Machine, 1769-1790 – Wolfgang von Kemplelen**
  - 第一台机械发声的语音器
  - 1791年，匈牙利科学家Wolfgang von Kempelen利用一系列精巧的风箱、弹簧片、风笛与共鸣箱制造出一个可以发出简单词汇的机器，取名为“讲话机”。


Video from the opening ceremony of Interspeech 2019 @ Graz, Austria



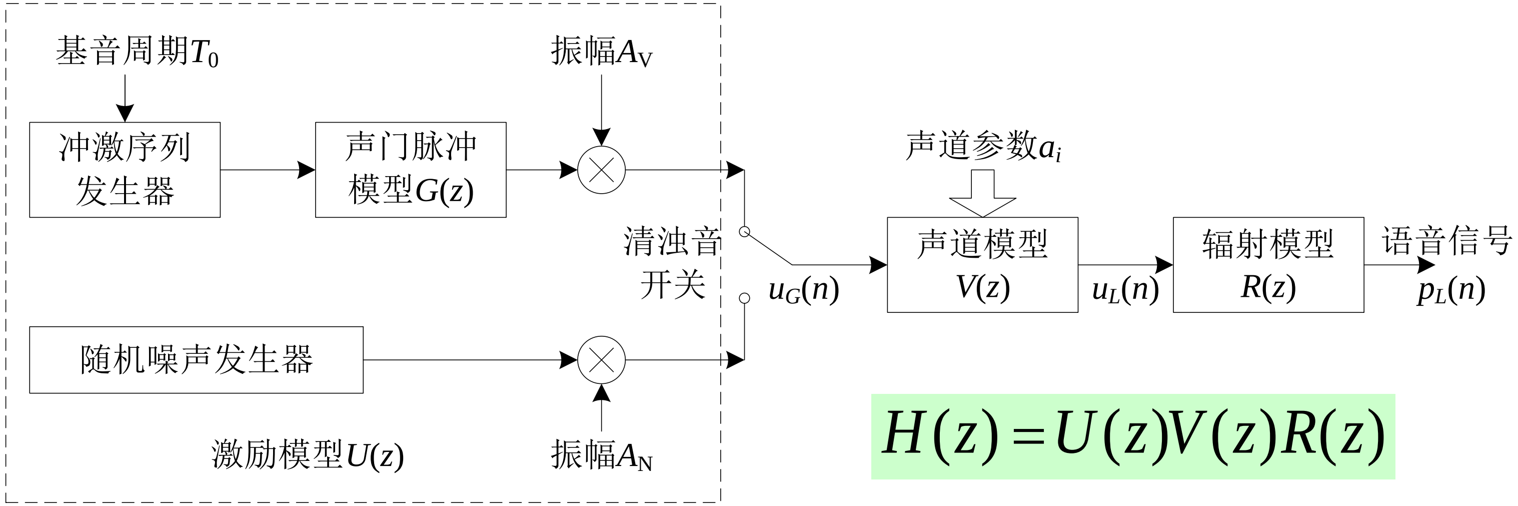
Speech Generation Model: 语音产生模型

- **Source-Filter Model:**
  源-滤波器模型
  - 语音是由气流激励声道，最后从嘴唇或鼻孔辐射出来而形成



Source-filter model and the corresponding spectra

- **The Block Diagram**
  - Excitation　激励模型：声源(准周期气流脉冲或白噪声)去激励声道
  - Vocal Tract 声道模型：对声音的调制、谐振作用；共振峰模型
  - Radiation　辐射模型：嘴唇或鼻孔的辐射效应



基音周期T_0

冲激序列发生器

声门脉冲模型$G(z)$

振幅A_V

清浊音开关

随机噪声发生器

振幅A_N

激励模型$U(z)$

$u_G(n)$

声道参数a_i

声道模型$V(z)$

$u_L(n)$

辐射模型$R(z)$

语音信号

$p_L(n)$

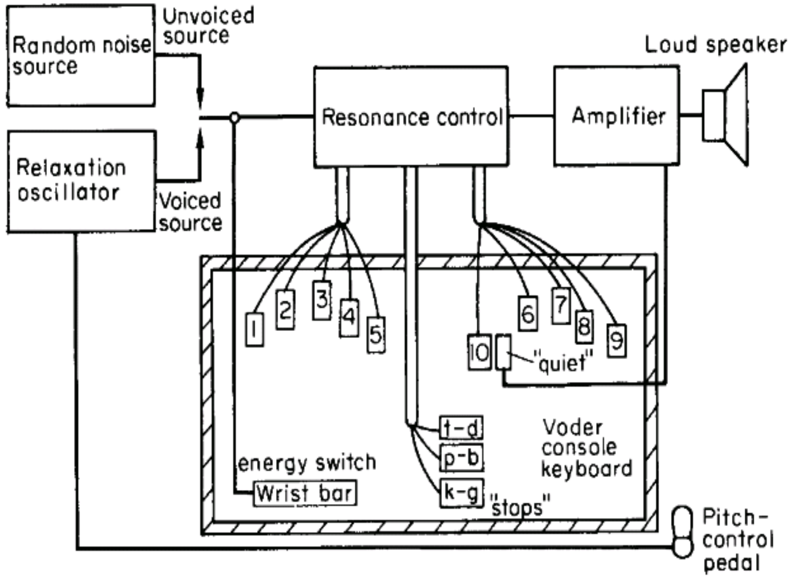
$$H(z)=U(z)V(z)R(z)$$

(reference: http://en.wikipedia.org/wiki/Source-filter_model_of_speech_production)

Predecessor of Speech Synthesizer

- **The Voder, 1939 – Homer Dudley, Bell Laboratory**
  - 1939年，Bell Lab的H. Dudley制作的语音合成器VODER（VOice DEmonstratoR）在纽约博览会上展出
  - 第一台电子语音合成器
    - 手腕控制声源开关
    - 脚踏板控制张弛振荡器以改变声调
    - 手指控制10个琴键，以控制带通滤波器（共振器）
    - 3个额外的琴键控制爆破音

http://www.ling.su.se/staff/hartmut/kemplne.htm



The operator needed a year’s practice just to master the keys!

Types of Speech Synthesis

- **Articulatory Synthesis, 发音器官参数合成**
  - Model movements of articulators and acoustics of vocal tract
  - 对人的发音过程进行直接模拟
- **Formant Synthesis, 共振峰合成, 声道模型参数合成**
  - Start with acoustics, create rules/filters to create each formant
  - 基于声道截面积函数或声道谐振特性合成语音
- **Concatenative Synthesis, 拼接式语音合成**
  - Use databases of stored speech to assemble new utterances
  - 基于语音数据库从中挑选语音单元合成语音
- **Parametric Synthesis, 参数合成**
  - Speech synthesis based on HMM model
  - Speech synthesis based on DNN related model
  - 基于HMM/DNN模型生成基频、频谱等声学特征参数，然后再使用参数合成器合成语音

Articulatory Synthesis

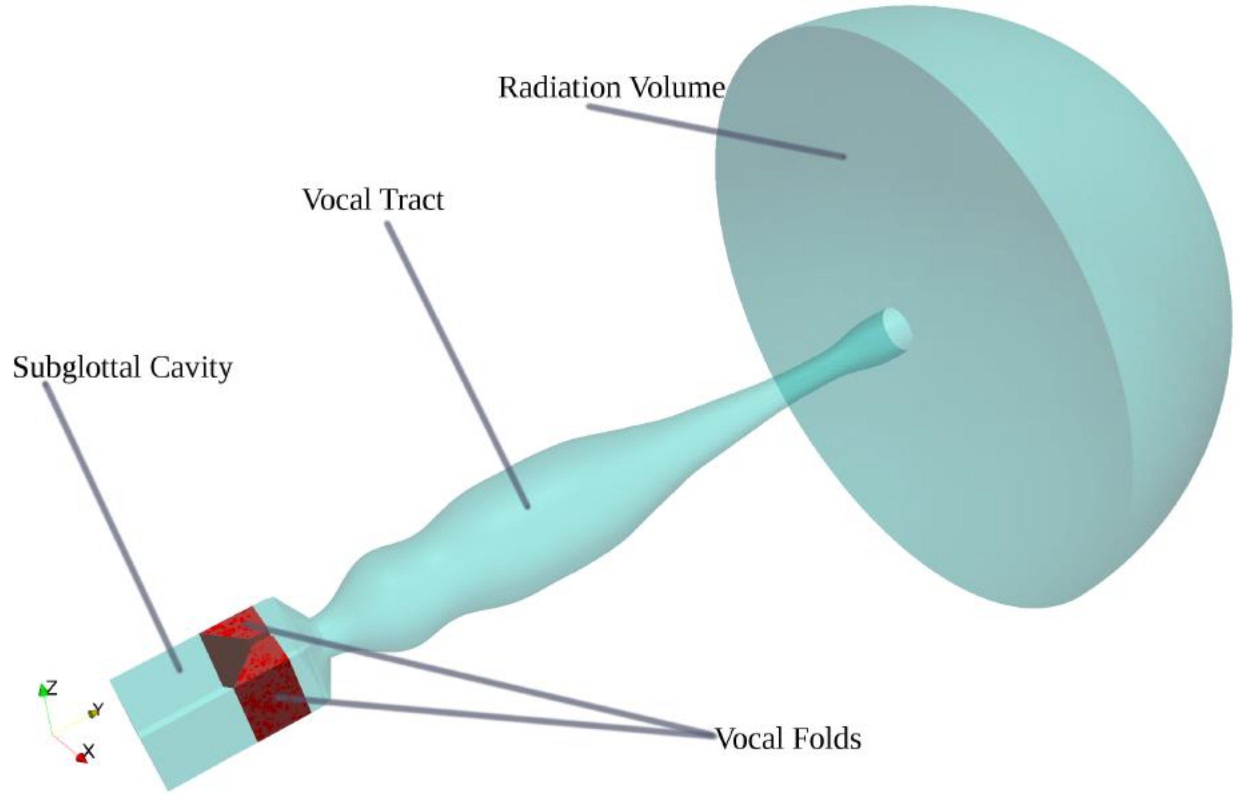


Figure 1: *Computational domain for the unified numerical production of vowel [i].*

(reference: N.C. Degirmenci, J. Jansson, J. Hoffman, et al., A unified numerical simulation of vowel production that comprises phonation and the emitted sound, 2017)

Articulatory Synthesis

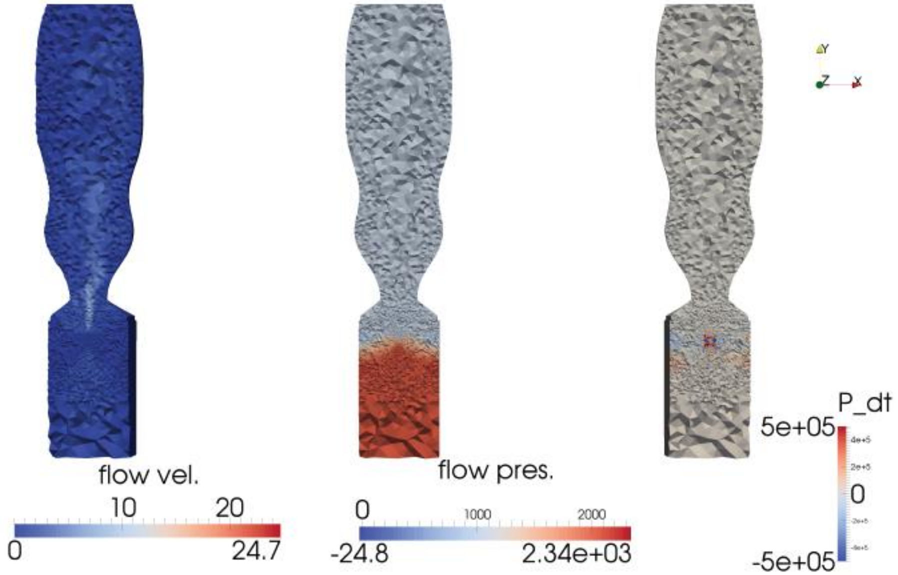


Figure 3: *Snapshot of the hydrodynamic velocity, hydrodynamic pressure and its derivative for unified simulations of vowel [i] at* $t = 0.055$.

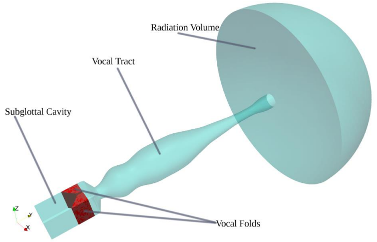


Figure 1: *Computational domain for the unified numerical production of vowel [i].*

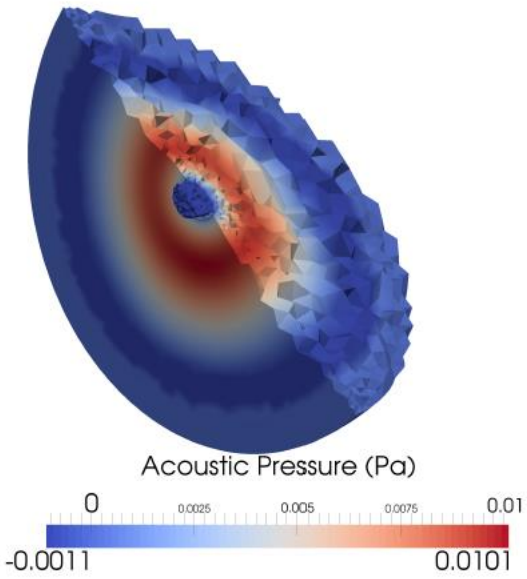


Figure 4: *Snapshot of the acoustic pressure for the unified simulation of vowel [i] at* $t = 0.055$.

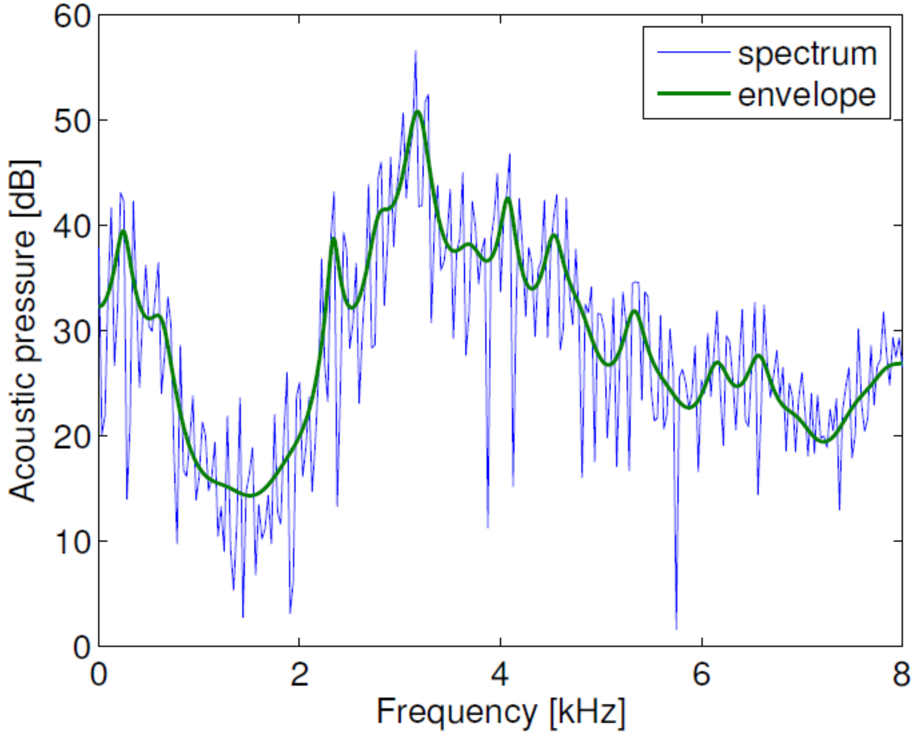
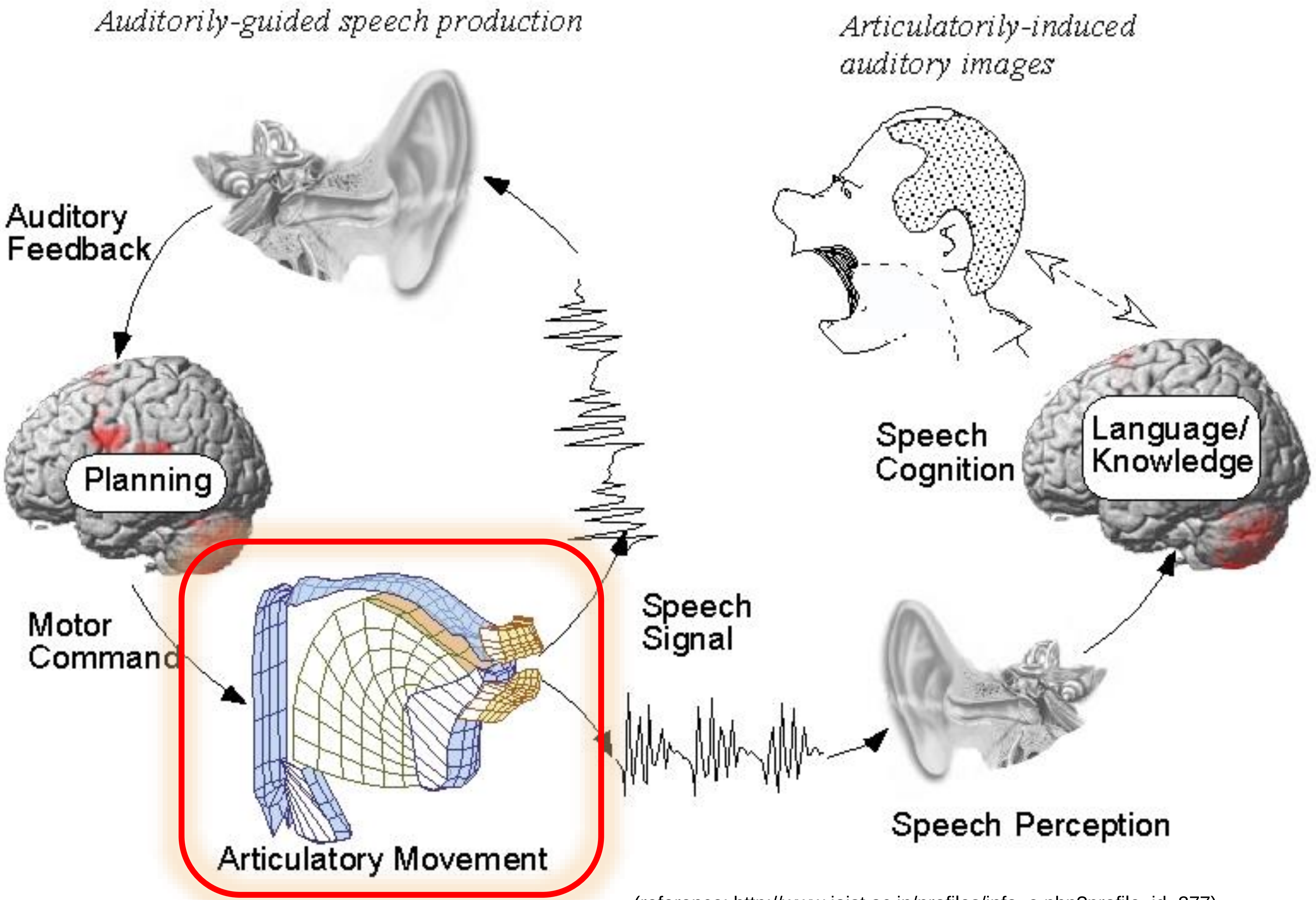


Figure 5: *Spectrum of the acoustic pressure at a point located at position* $(-0.0036, 0.2480, 0.0151)$ *outside the mouth. The formants of vowel [i] can be clearly appreciated.*

Articulatory Synthesis



(reference: http://www.jaist.ac.jp/profiles/info_e.php?profile_id=277)

Articulatory Synthesis

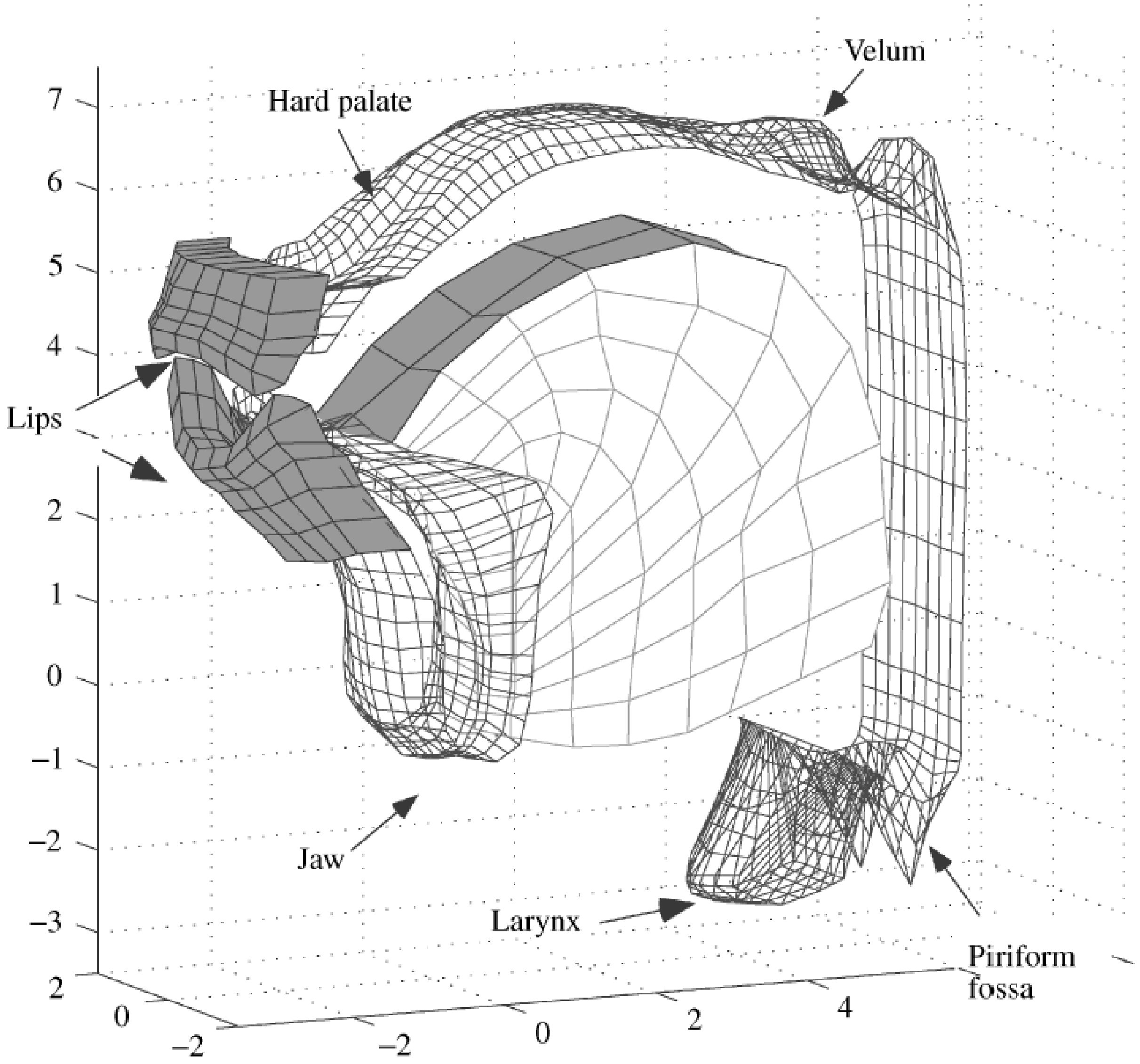


Figure 1. Configuration of the physiological articulatory model.

(reference: Jianwu Dang, Kiyoshi Honda, Estimation of vocal tract shapes from speech sounds with a physiological articulatory model, 2002)

Formant Synthesis

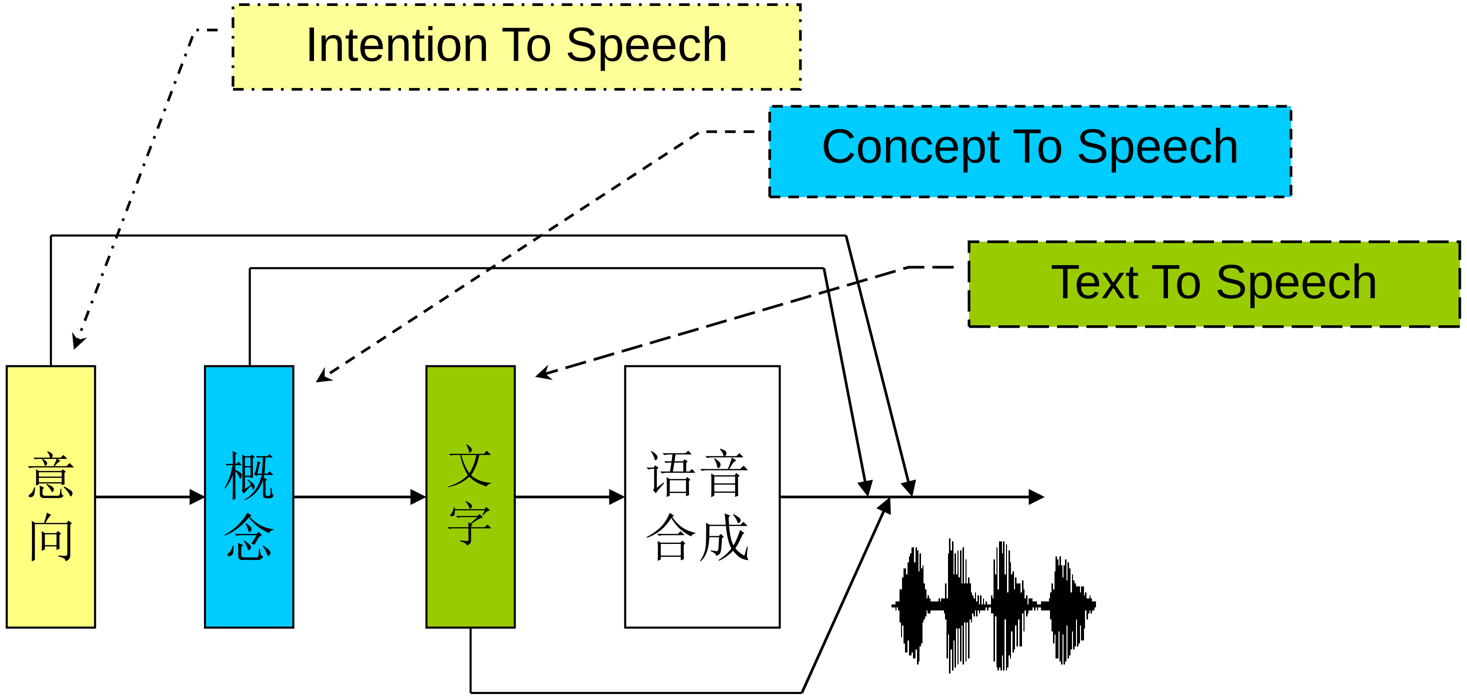
- **Were the most common commercial systems while computers were relatively underpowered.**

- 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1983 DECtalk system

文语转换

Text-to-Speech, TTS

Speech Generation

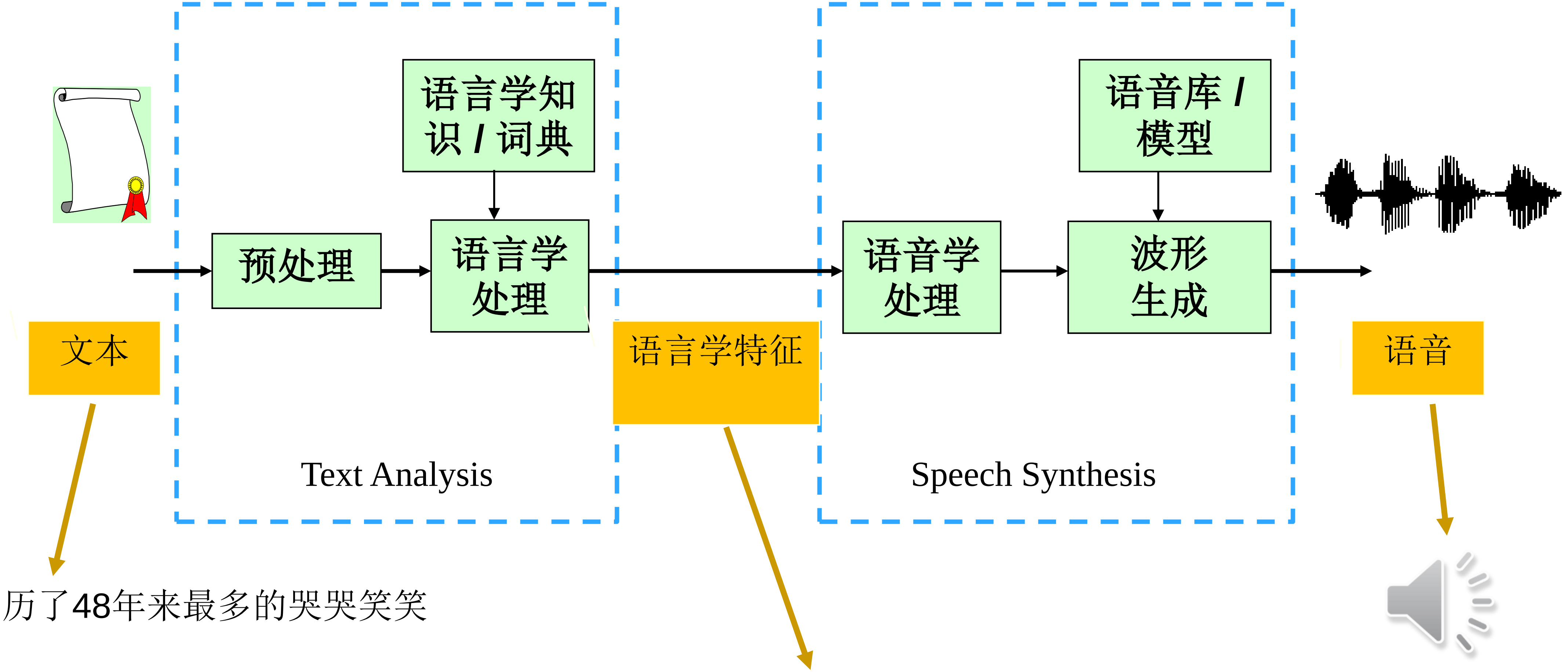


TTS: Text-to-Speech

- 文语转换
  - 文语转换（Text-to-Speech，TTS）：把计算机内的文本转换成连续自然的语声流。
  - Text-to-Speech synthesis involves the generation of a speech signal from the input text.

Architecture of a TTS System

- **TTS的系统结构**



语言学知识 / 词典

预处理

语言学处理

语音库 / 模型

语音学处理

波形生成

文本

语言学特征

语音

Text Analysis

Speech Synthesis

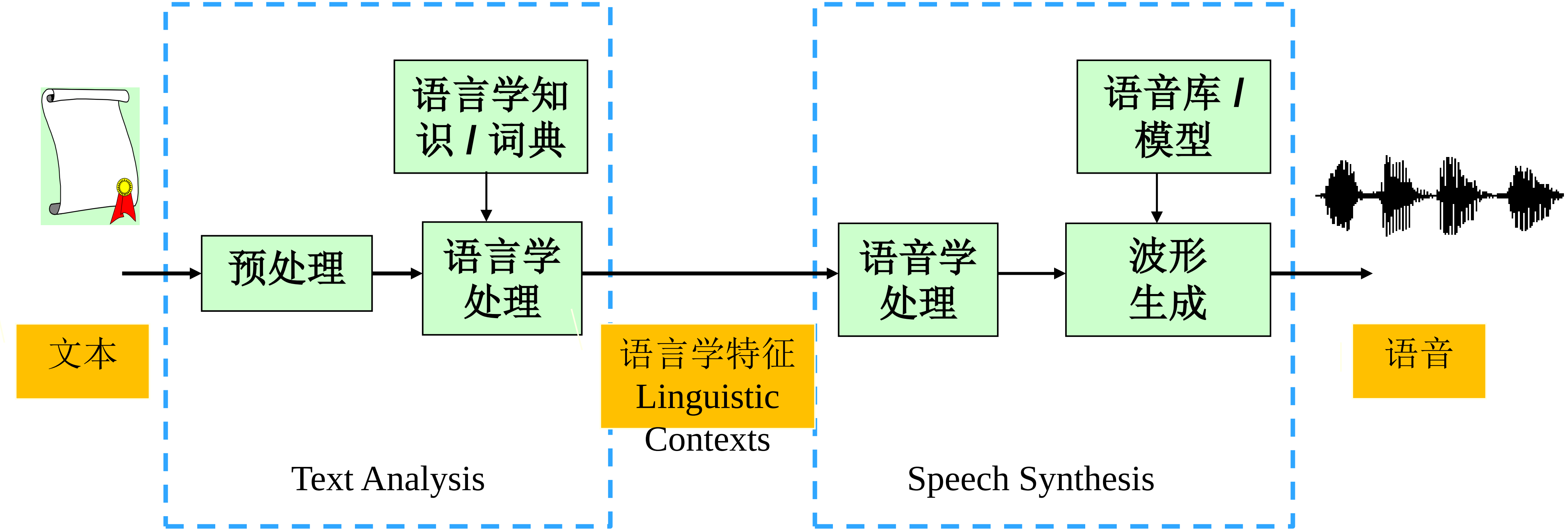
他说他经历了48年来最多的哭哭笑笑

/ 他说 / 他|经历了|四十八|年来 / 最多的|哭哭|笑笑 /

ta1 shuo1 ta1 jing1 li4 le5 si4 shi2 ba1 nian2 lai2
zui4 duo1 de5 ku1 ku1 xiao4 xiao4

Typical Flow of a TTS System

- **TTS的系统结构**



- **Text Analysis (frontend)**
  - Text normalization
  - Word segmentation
  - Part-of-speech (POS) tagging
  - Grapheme to phoneme (G2P)
  - Prosody prediction

- **Speech Synthesis (backend)**
  - Two Approaches
    - Unit Selection Synthesis
    - Statistical Parametric Synthesis

Linguistic Features / Linguistic Contexts

- **影响发音的上下文信息**
  - 当前音节信息
    - 当前音节读音
    - 声母类型
    - 韵母类型
    - 声调信息
  - 前后音节信息
    - 相邻前音节的韵母类型
    - 相邻前音节的声调类型
    - 相邻后音节的声母类型
    - 相邻后音节的声调类型
  - 在韵律层级结构中的位置信息
    - 语句
    - 韵律短语
    - 韵律词
  - ……

/ 他说 / 他|经历了|四十八|年来 / 最多的|哭哭|笑笑 /

ta1 shuo1 ta1 jing1 li4 le5 si4 shi2 ba1 nian2 lai2 zui4 duo1 de5 ku1 ku1 xiao4 xiao4

Modeling Unit

- **Definition of unit**
  - Diphone
    - Middle of one phone to middle of next
    - Why? Middle of phone is steady state.
  - Triphone
    - Middle of previous phone to whole current phone to middle of next phone
  - Initial/Final
  - Syllable
  - Word
  - Even larger units

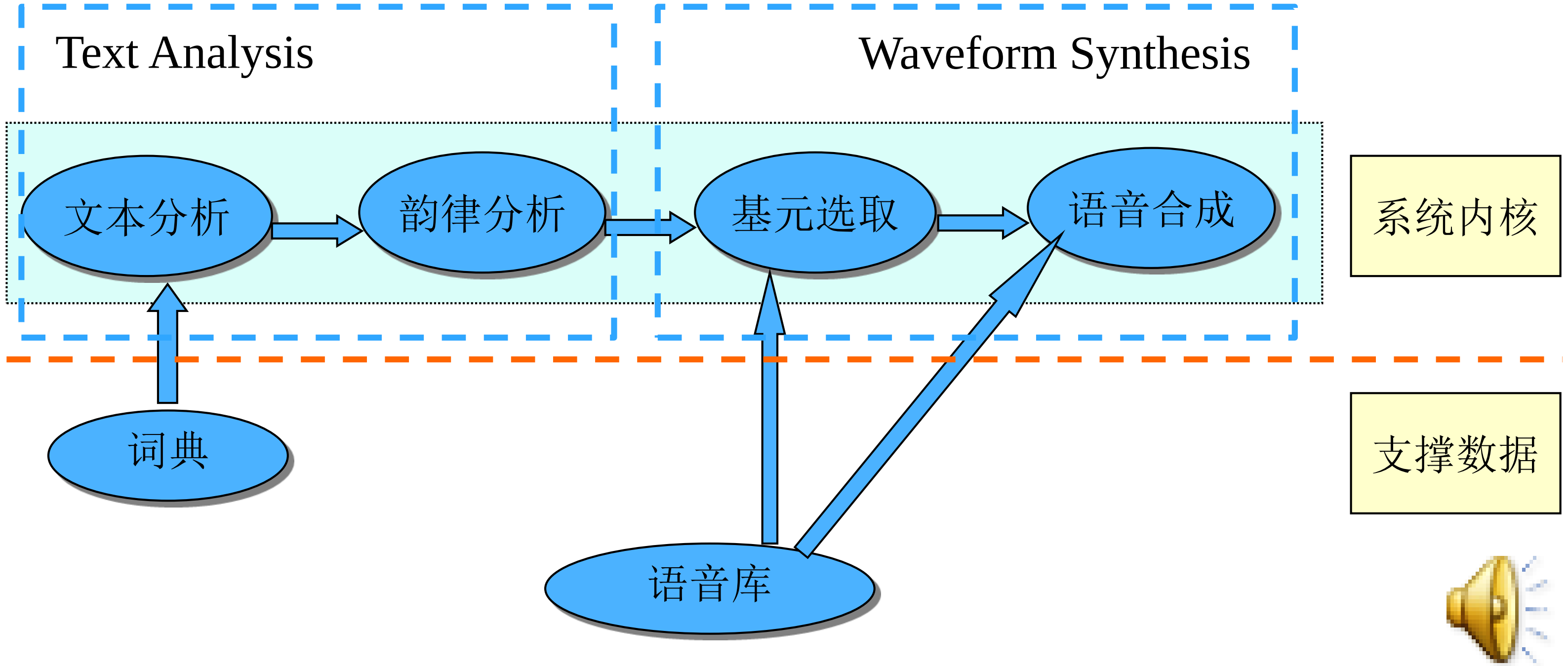
Speech Database and Synthesis Approaches

- **Speech database / Speech corpus**
  - Record 10 hours or more
  - To have multiple copies of each unit, which each unit corresponding to different context of linguistic features

- **Unit Selection Synthesis**
  - Use search to find best sequence of units based on "linguistic features"
- **Statistical Parametric Synthesis**
  - Find the mapping function between "linguistic features" and "acoustic features", where "acoustic features" are extracted from the speech units

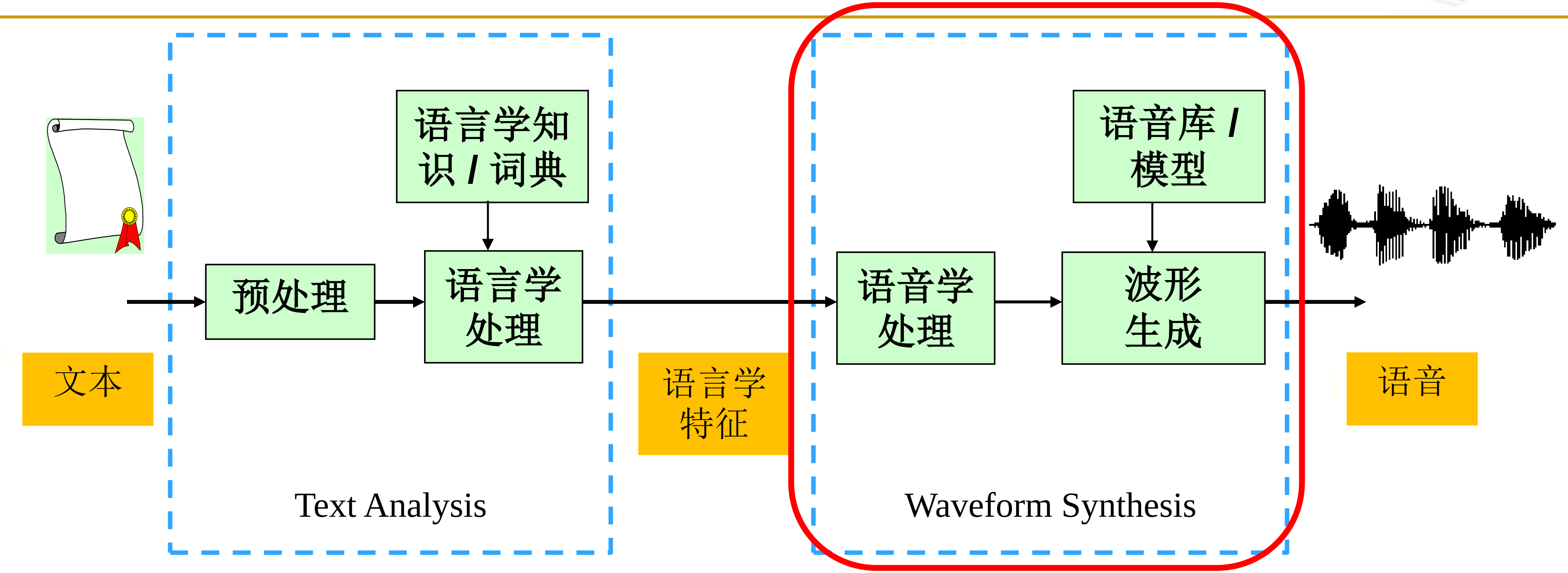
Concatenative Speech Synthesis

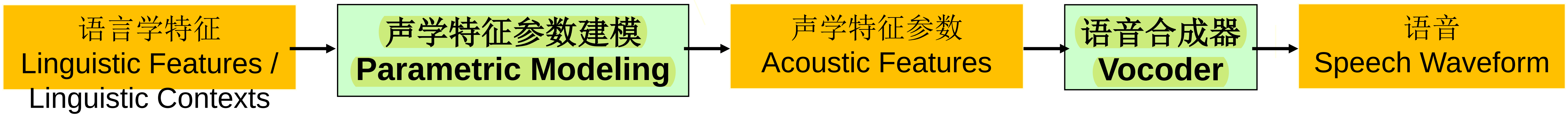
- 拼接式语音合成
  - 直接把语音波形数据库中的波形拼接在一起，输出连续语流。
  - 利用基元选取算法（Unit Selection Algorithm）基于语言学特征的上下文信息从语音库中选择“最合适”的语音单元。这些语音单元取自自然语音的词或句子，隐含了声调、重音等细微特性。



他五点钟值班。

Parametric Speech Synthesis



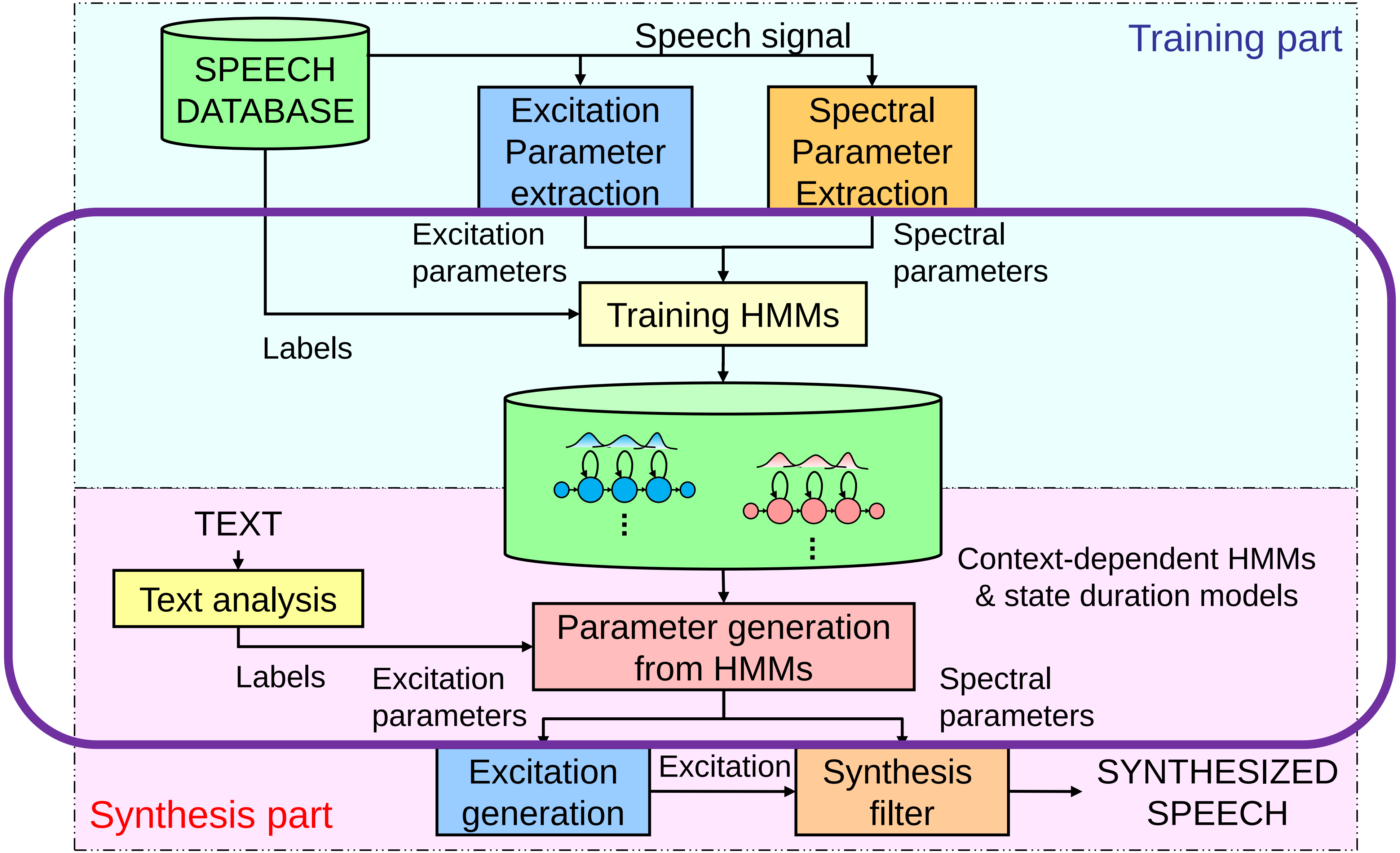


HMM
LSTM

Duration Parameters: Duration
Excitation Parameters: Pitch
Spectral Parameters: Spectrum

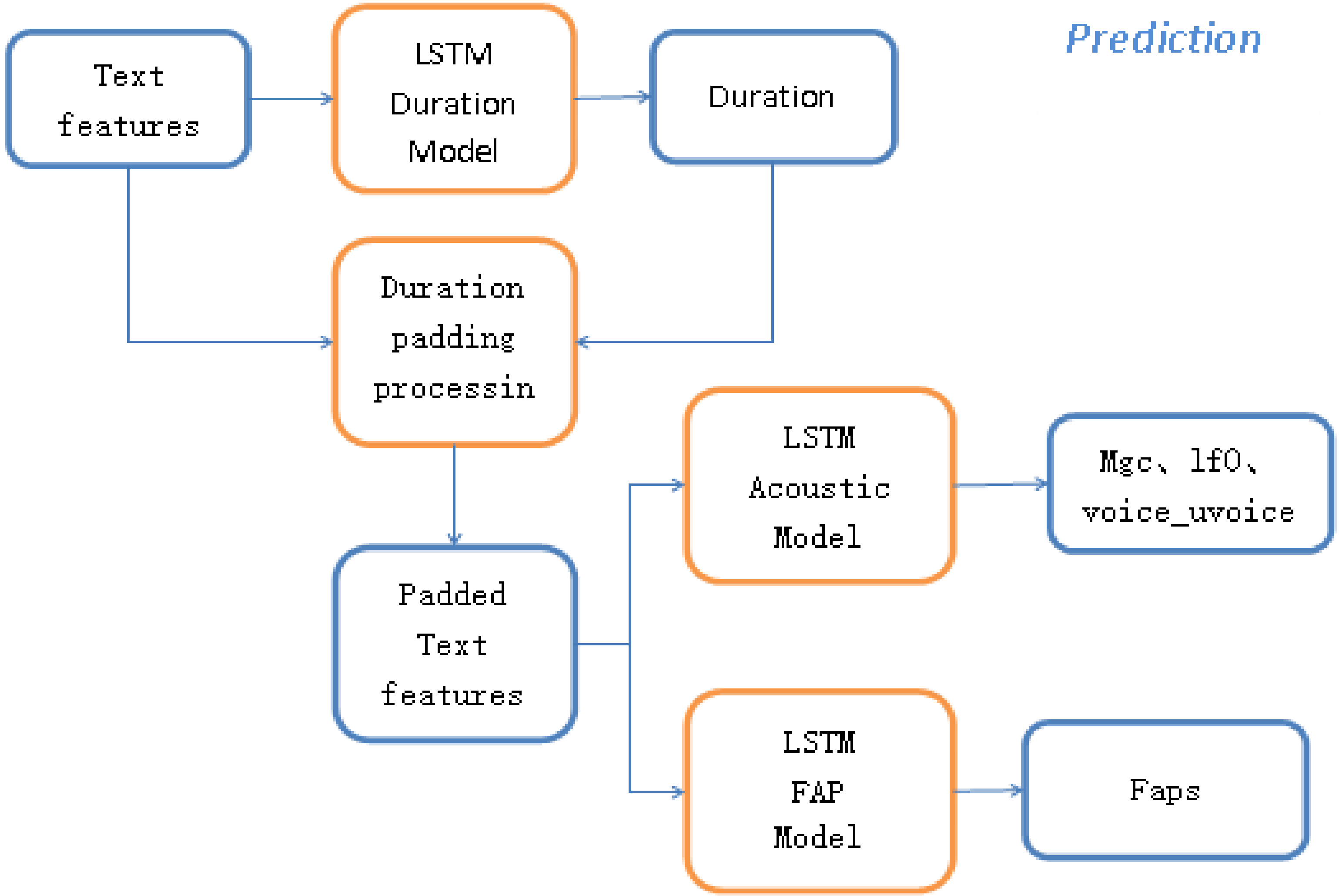
SPTK
STRAIGHT

HMM based Parametric Speech Synthesis



(Fig. from HTS)

LSTM based Parametric Speech Synthesis

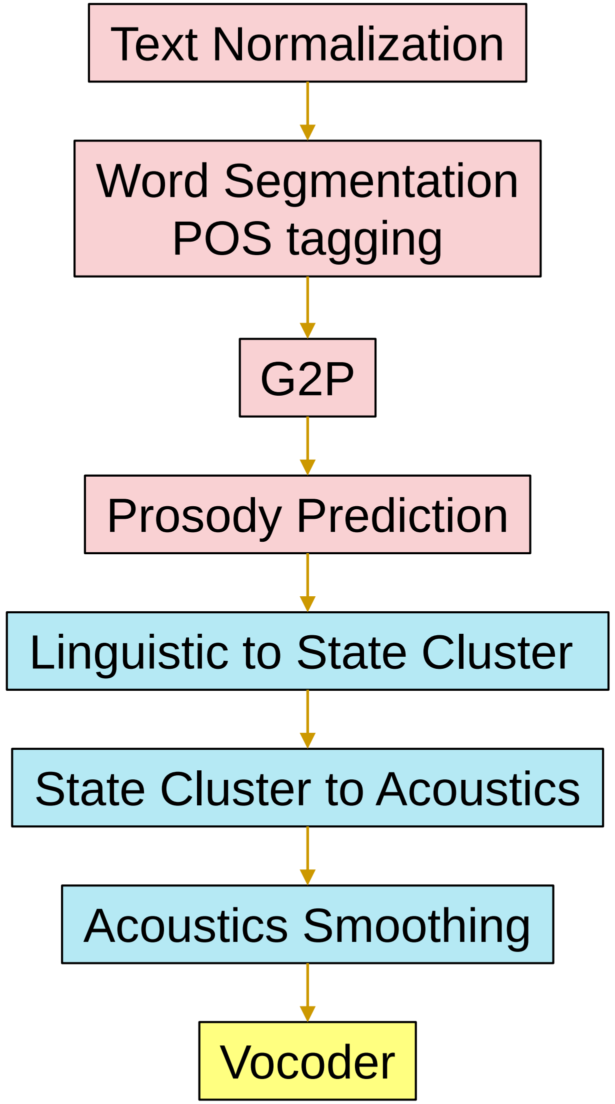


Prediction

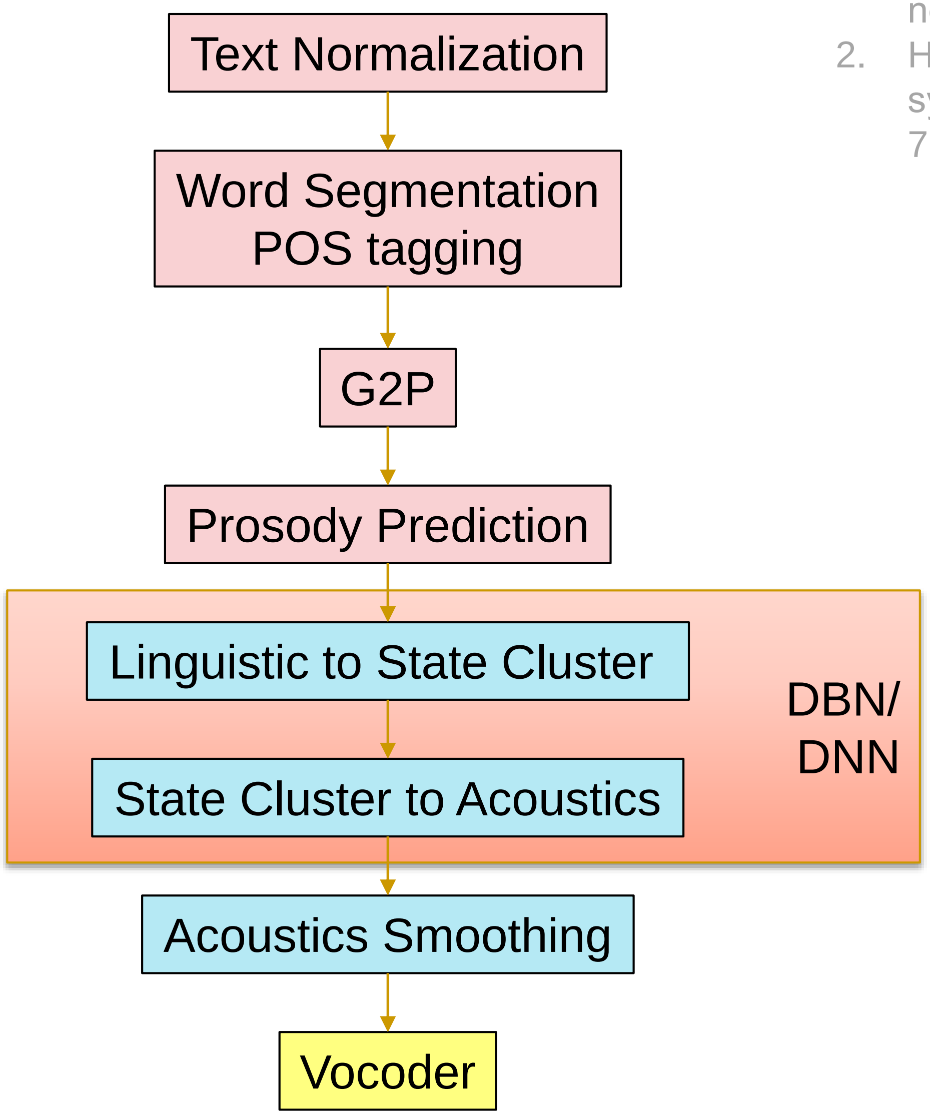
(Fig. from Li Xu)

Beyond Parametric TTS

Beyond Parametric TTS

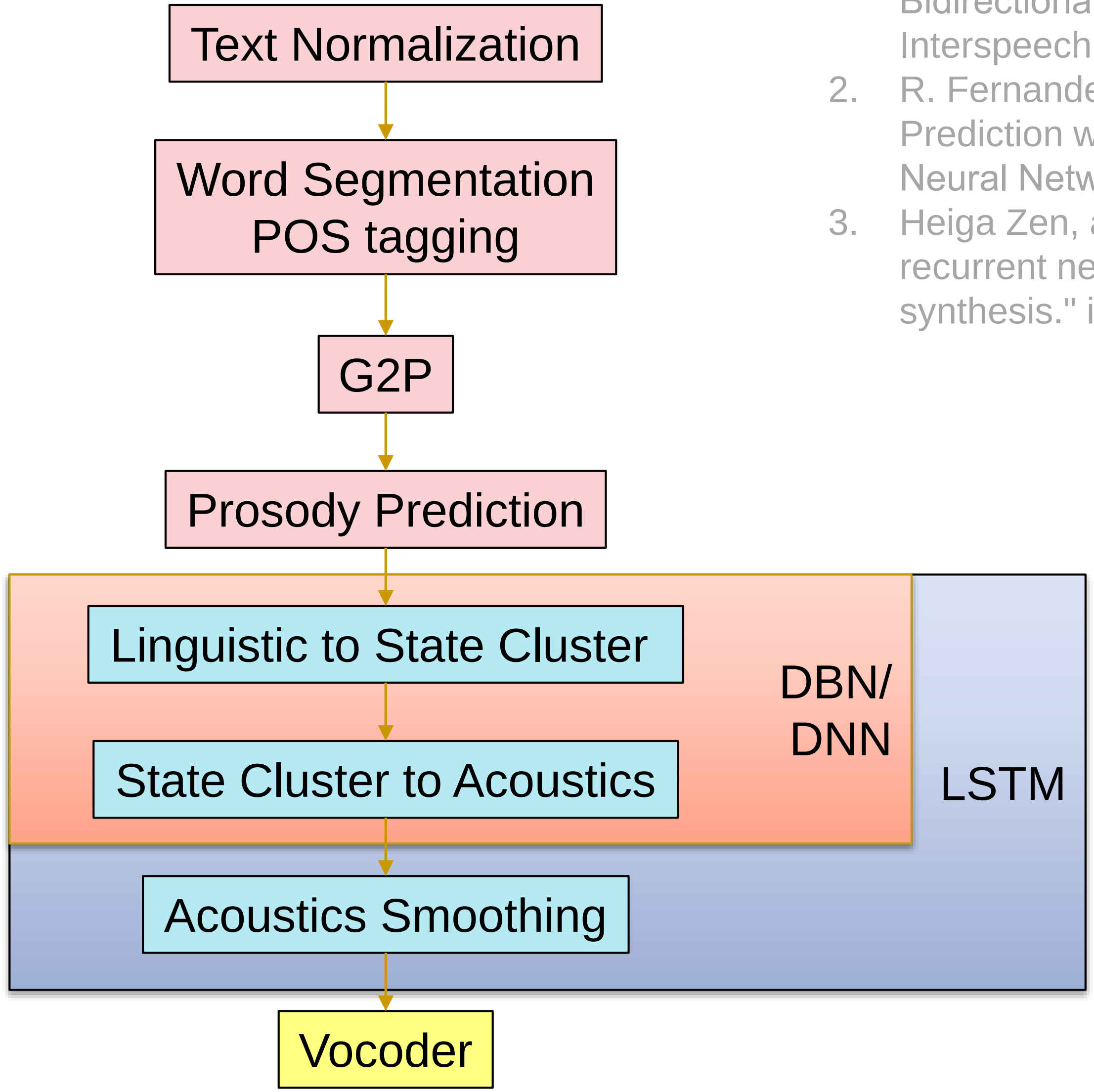


Beyond Parametric TTS


Text Normalization
Word Segmentation POS tagging
G2P
Prosody Prediction
Linguistic to State Cluster
State Cluster to Acoustics
DBN/ DNN
Acoustics Smoothing
Vocoder

1. Shiyin Kang, Xiaojun Qian, and Helen Meng, “Multi-distribution deep belief network for speech synthesis,” in Proc. ICASSP, 2013, pp. 8012–8016.
2. H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP, 2013, pp. 7962–7966.

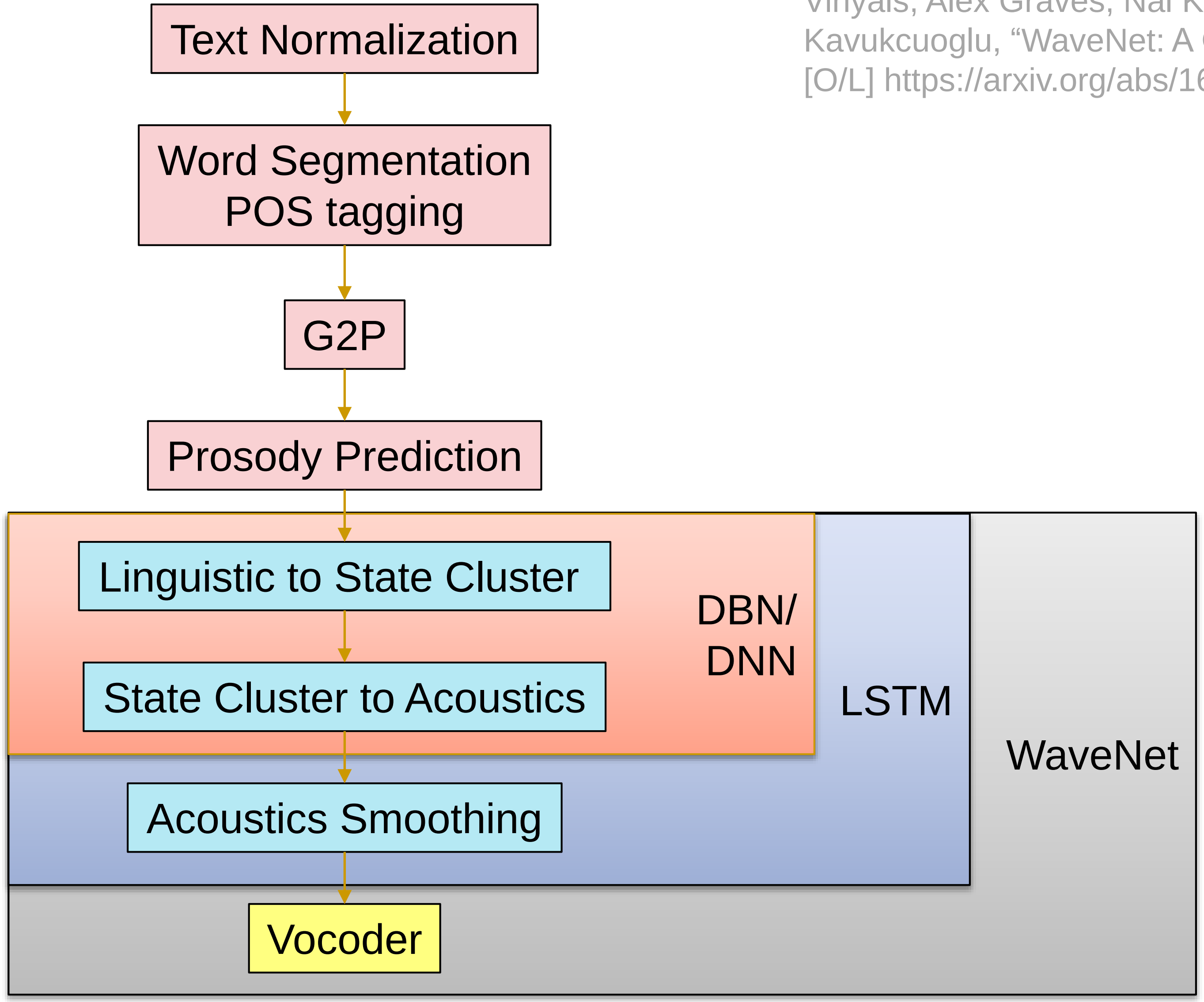
Beyond Parametric TTS



1. Y.C. Fan, Y. Qian, F.L. Xie and F.K. Soong, “TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks”, in Proc. Interspeech, 2014.
2. R. Fernandez, A. Rendel, B. Ramabhadran, R. Hoory, “Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional Deep Recurrent Neural Networks”, in Proc. Interspeech, 2014.
3. Heiga Zen, and Haşim Sak. "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis." in Proc. ICASSP, 2015.

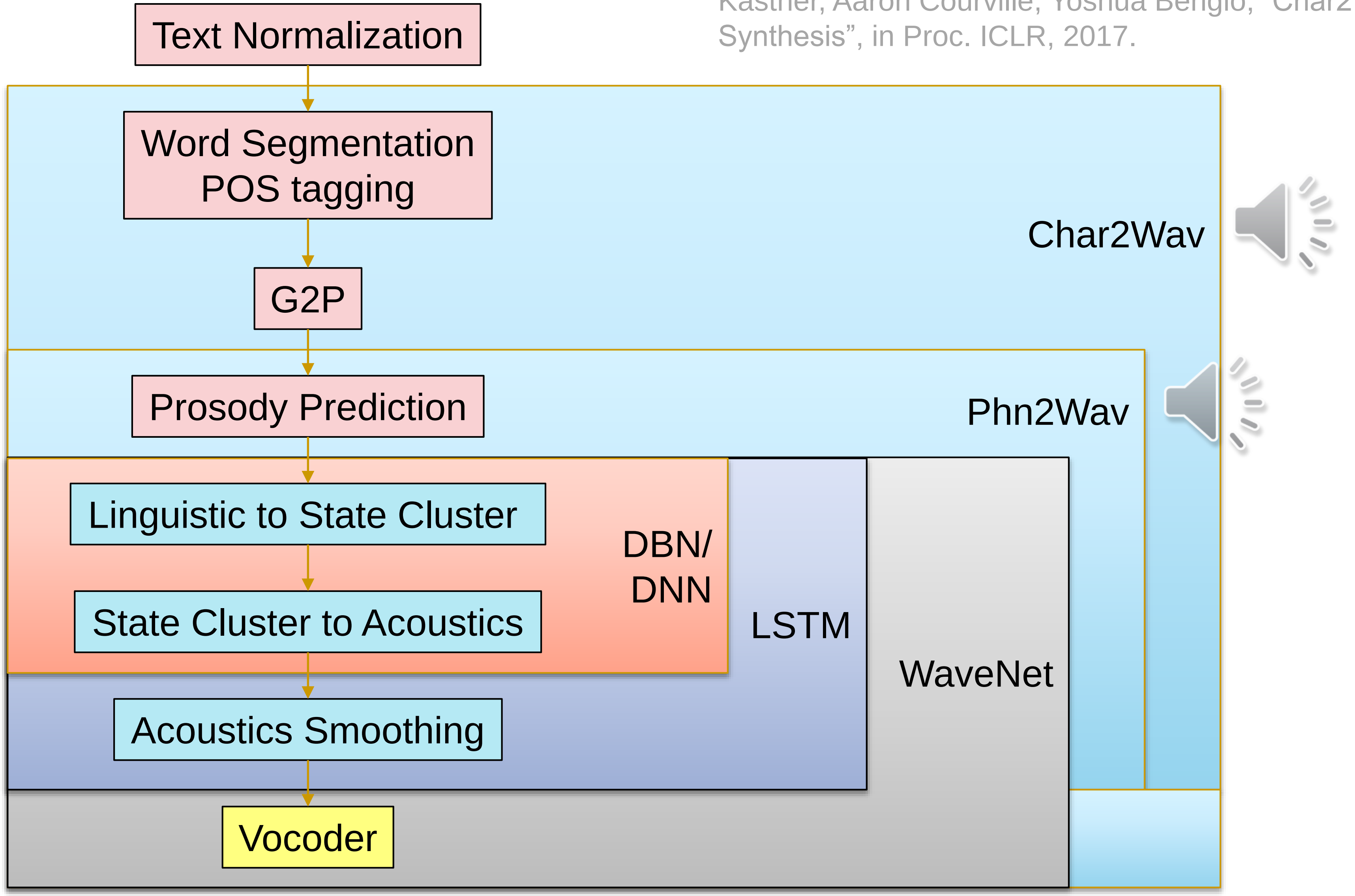
Beyond Parametric TTS

1. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio”, [O/L] https://arxiv.org/abs/1609.03499, 2016.



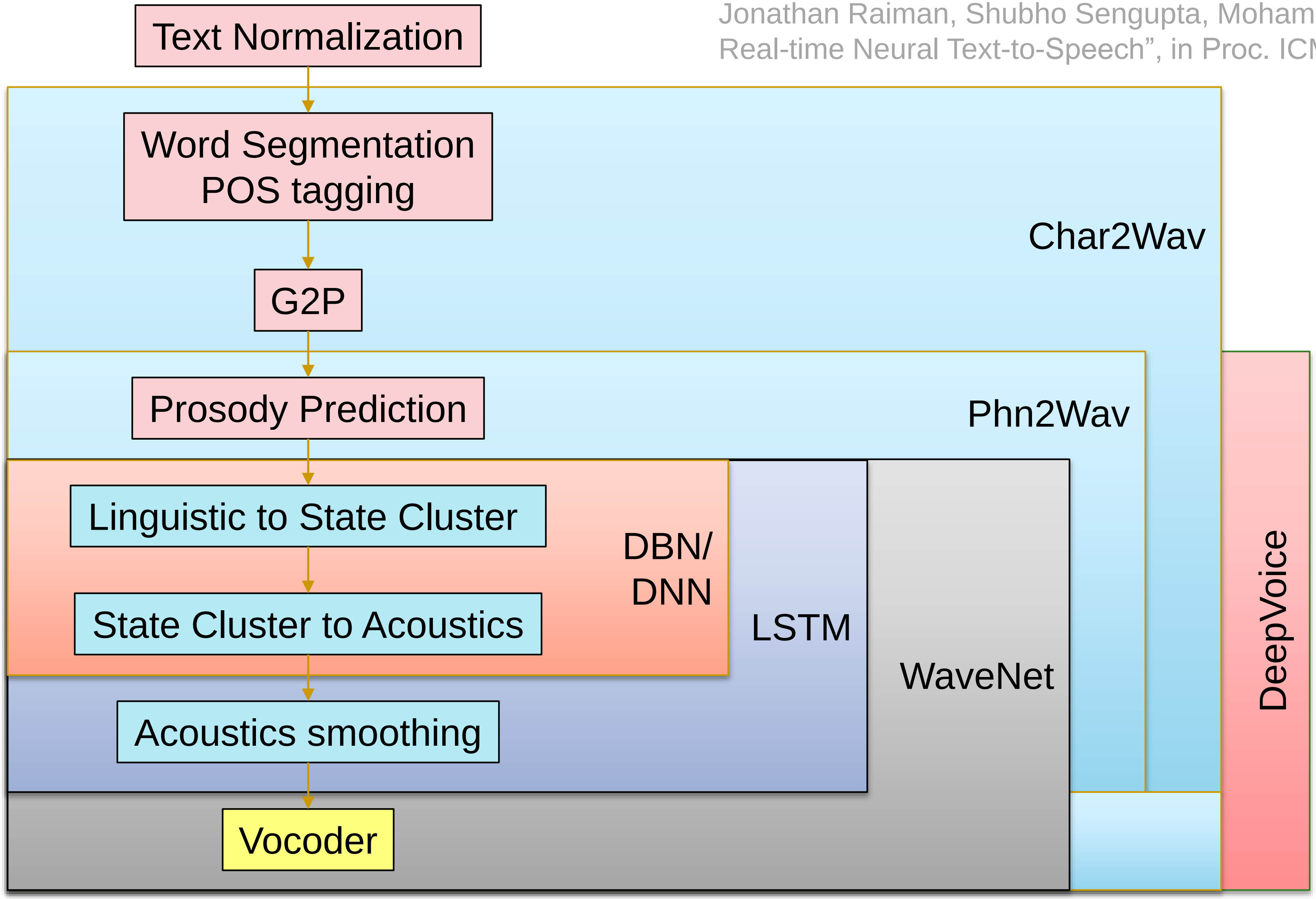
Beyond Parametric TTS

1. Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, Yoshua Bengio, “Char2Wav: End-to-End Speech Synthesis”, in Proc. ICLR, 2017.



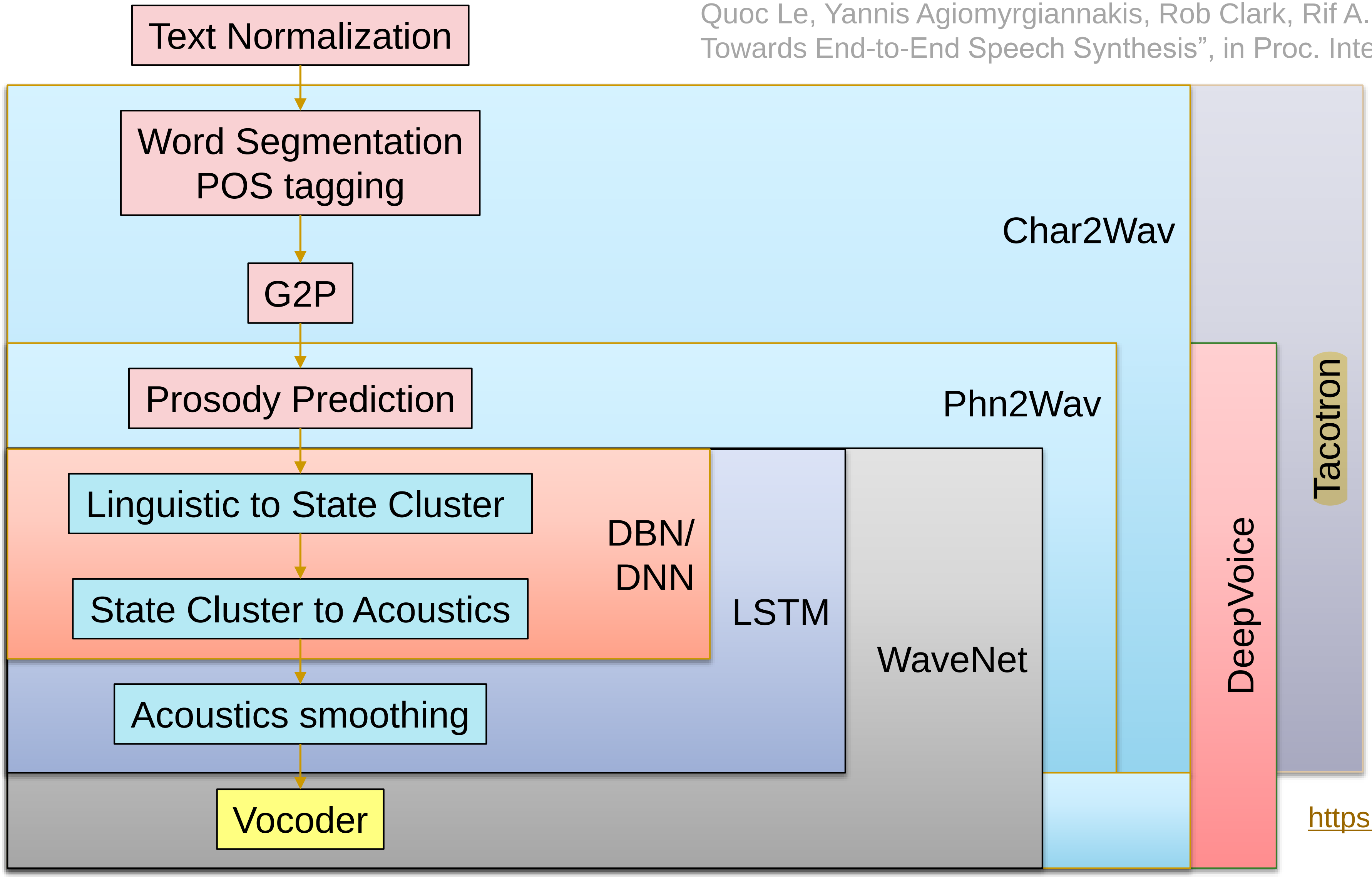
Beyond Parametric TTS

1. Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi, “Deep Voice: Real-time Neural Text-to-Speech”, in Proc. ICML, 2017.



Beyond Parametric TTS

1. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis", in Proc. Interspeech, 2017.



https://google.github.io/tacotron/

Applications of TTS

- Conversational agents that conduct dialogues with people, e.g., call center
- Non-conversational application that speak to people, such as in devices that read out for blind, information kiosks, or in video games or children’s toys
- Speaking for suffers of neurological disorders, such as astrophysicist Stephen Hawking

QTOONS
www.count5.com
CALL CENTE
THAT OBJECTION WASN'T COVERED IN MY SALES TRAINING... SO, LET'S JUST SKIP IT AND STICK TO MY SCRIPT, OK?
CHAD
Five hundred twenty-six.
526

TTS Demos

- Festival
  - http://www-2.cs.cmu.edu/~awb/festival_demos/index.html
- Cepstral
  - http://www.cepstral.com/cgi-bin/demos/general
- IBM
  - http://www-306.ibm.com/software/pervasive/tech/demos/tts.shtml
- AT&T
  - http://www.research.att.com/~ttsweb/tts/demo.php
- iFlyTek, USTC
  - http://www.iflytek.com/TtsDemo/interPhonicShow.aspx
- Crystal, Tsinghua, CUHK
  - http://www.se.cuhk.edu.hk/crystal
- DeepVoice, Baidu
  - http://research.baidu.com/Blog/index-view?id=91
- Tacotron, Google
  - https://google.github.io/tacotron/
- WaveNet, DeepMind
  - https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

拼接式语音合成及其模块划分

CONCATENATIVE TTS & IT'S MODULE DIVISION

Module Division of Concatenative TTS

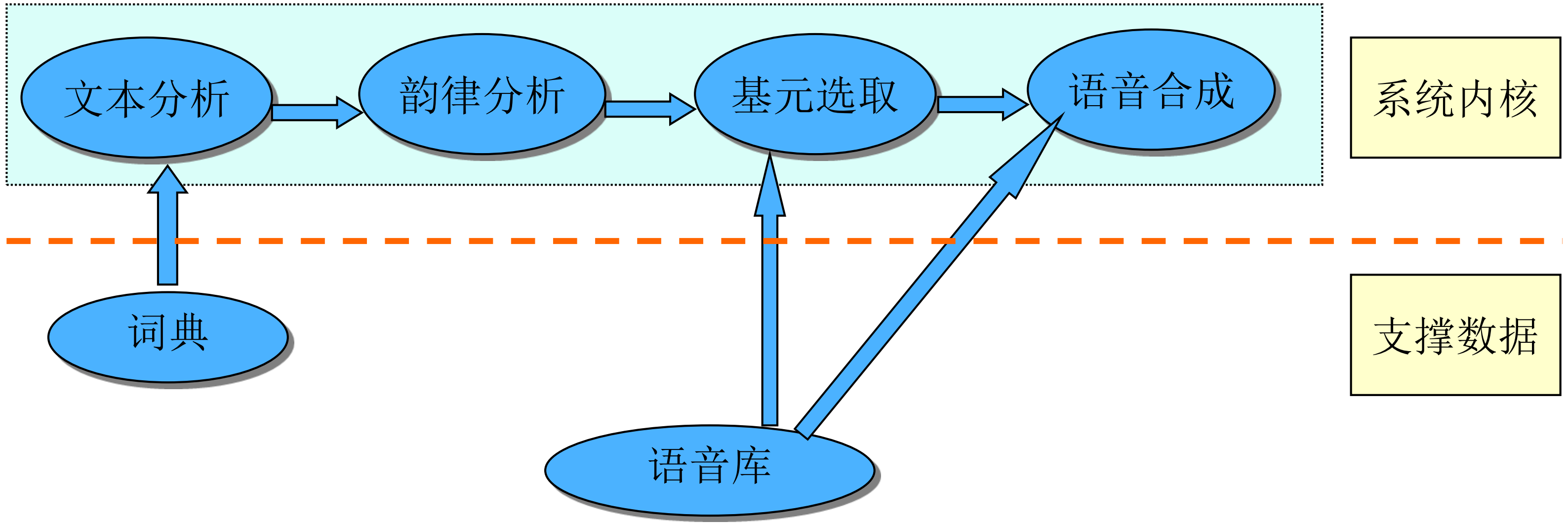
- **TTS的模块划分**

文本规范化
语法分析
语义分析
词法分词

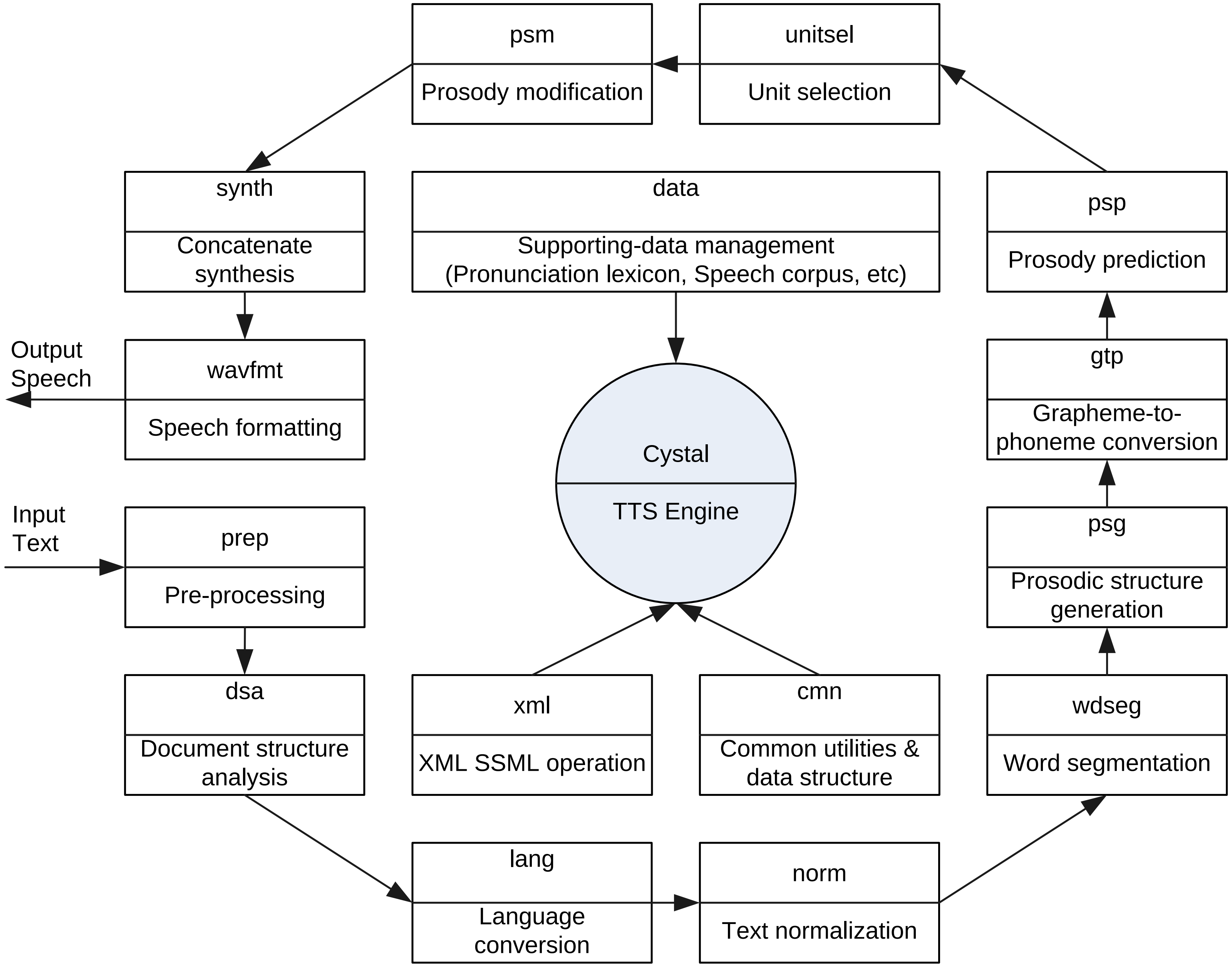
韵律结构分析
协同发音分析
字音转换
韵律预测
(轻重缓急)

根据读音及韵律结构、韵律特征信息从语音库中选取合适的语音基元

波形编辑、修改
拼接合成



Crystal: An Example



SSML: Speech Synthesis Markup Language

- 语音合成标记语言

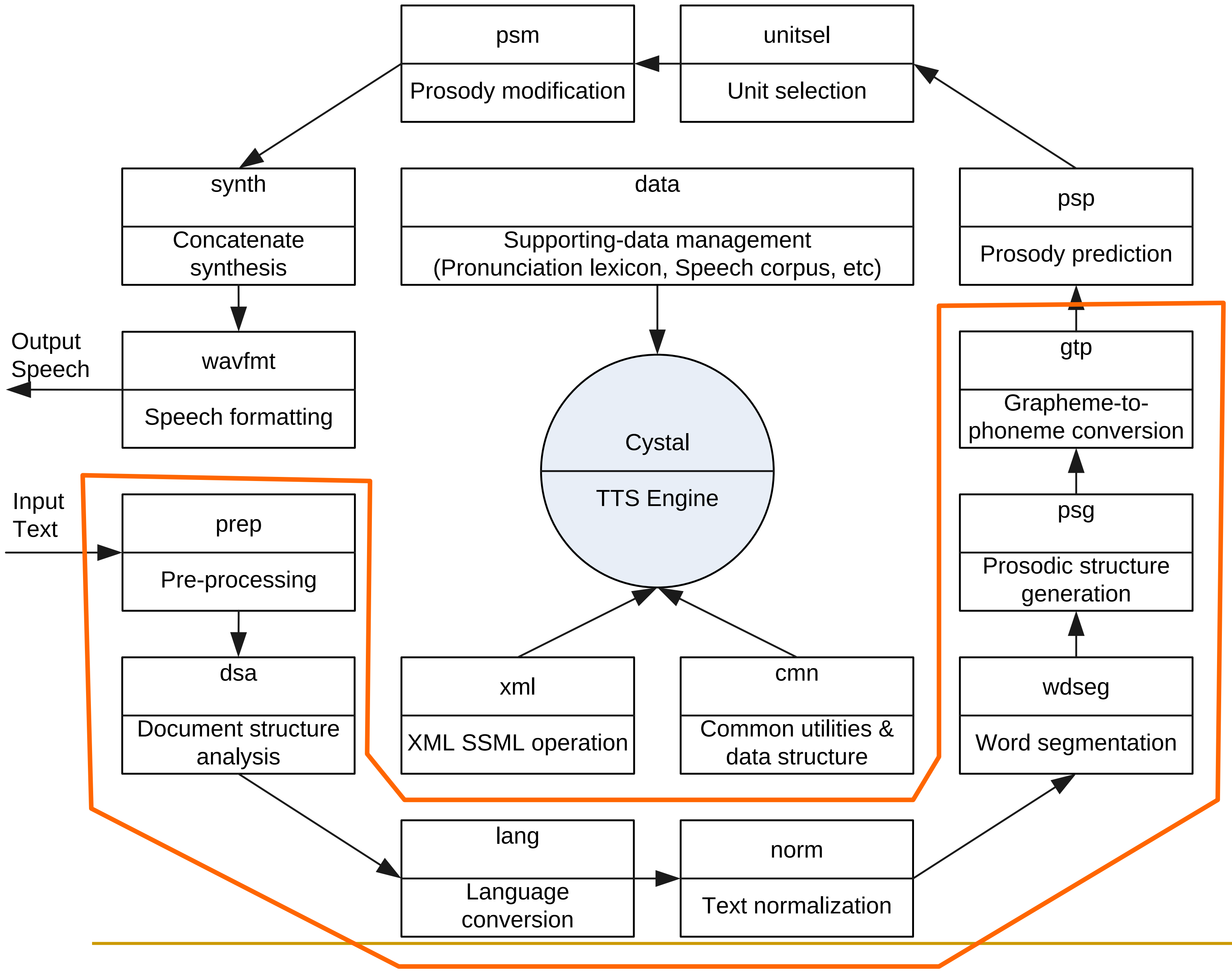
```
<?xml version="1.0" encoding="UTF-16"?>
<speak version="1.1"  xml:lang="zh-cmn">
<p>
    <s>
        <w role="t">
            <unit uid="jin1:0">
                <prosody duration="295ms">
                    <phoneme alphabet="pinyin" ph="jin1">今</phoneme>
                </prosody>
            </unit>
            <unit uid="tian1:2">天</unit>
        </w>
        <break strength=“medium" time=“100ms" />
        天气很好。</s>
    <s> 明天去动物园。</s>
    <s> 同一个笔画的字，依部首来排。</s>
</p>
</speak>
```

The data streaming interface between all modules of Crystal TTS engine.

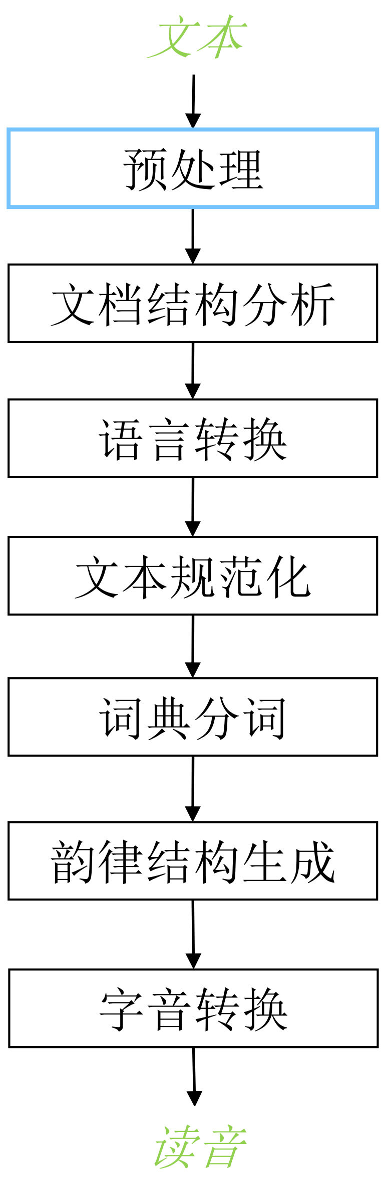
文本分析

TEXT ANALYSIS

Crystal: An Example



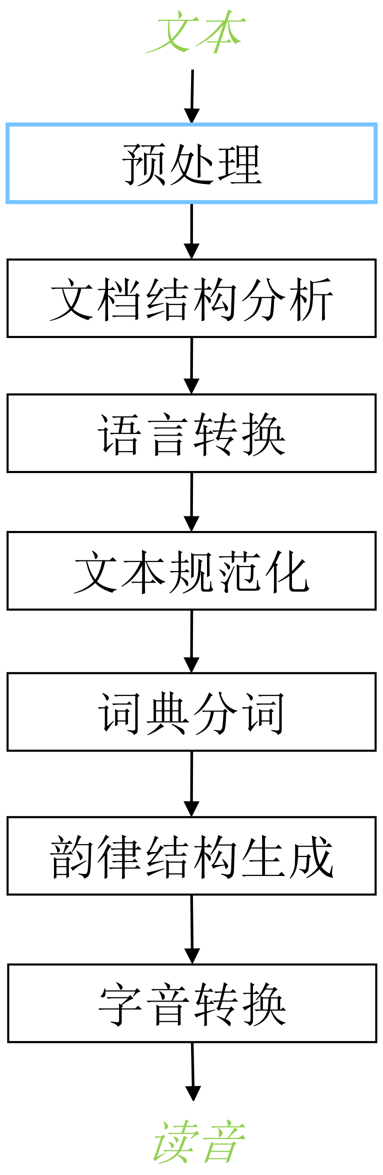
Pre-processing



- **To convert input text from any encodings to 16-bit Unicode Transformation Format (UTF-16)**
  - ❑ 编码分析
    - GB2312、GBK、BIG-5、UTF-8、UTF-16等
  - ❑ 编码转换
    - 将文本输入统一转换成UTF-16编码

- **To construct well-formed SSML document with proper header and elements**
  - ❑ SSML标注格式分析
  - ❑ 将预处理的结果转换成SSML格式化文档输出

Pre-processing

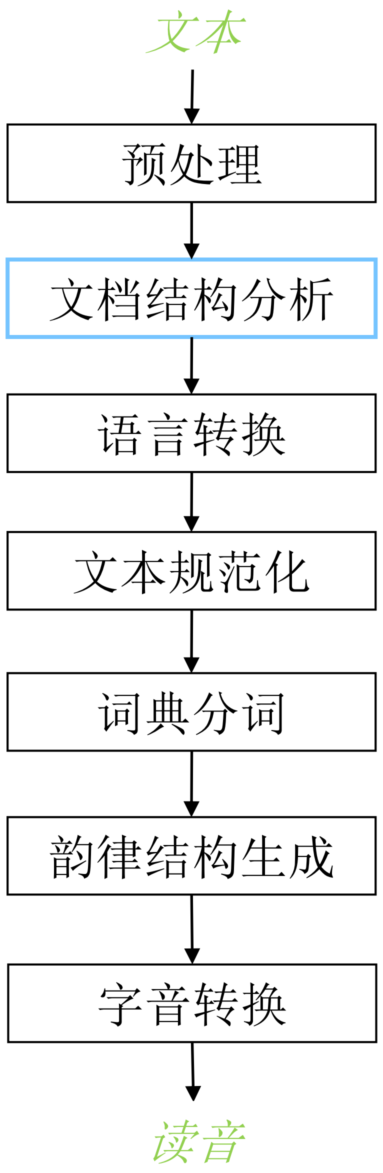


- **Original Input**
  - Pure text
  - Text with partial SSML tag(s)
  - Text with full SSML tag(s)
  - Example
    - 我说<w>道哥</w>，你还欠我HK$10,000.00呢！ 有冇搞錯！

- **Module Output**

```
<?xml version="1.0" encoding="UTF-16" ?>
<speak version="1.1" xml:lang="zh-cmn">
我说<w>道哥</w>，你还欠我HKD$10,000.00呢！
有冇搞錯！
</speak>
```

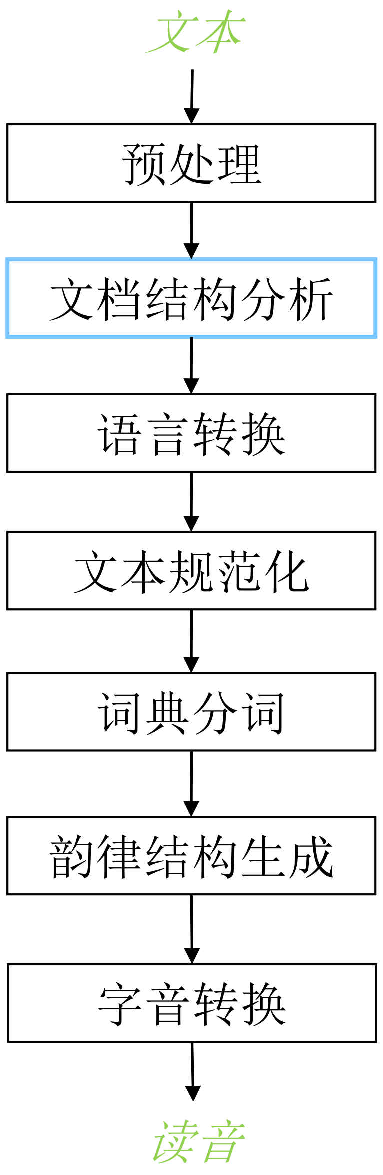
Document Structure Analysis



- **To segment input document into paragraphs and sentences**
  - 文档结构分析
    - 将输入的文本切分为段落和语句

- **To detect the natural language of the written content for each character piece**
  - (书写)语言检测
    - 检测每个文字片段的书写语言（如：简体中文、繁体中文）

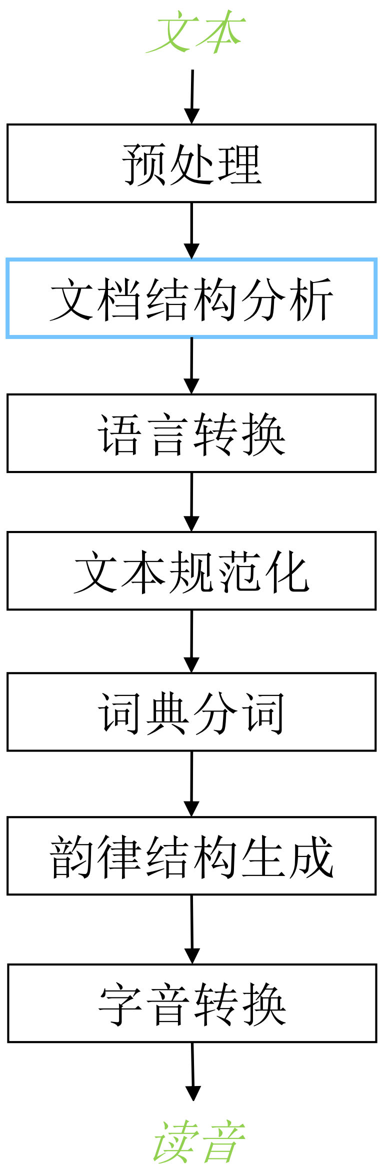
Document Structure Analysis



Regular Expression for Disambiguation

- **Punctuations are usually served as the signature of sentence delimiter**
  - Comma: “,”
  - Full-stop: “.”
- **However, some of them may also appear in many special types of input data constructs (*special constructs*)**
  - URLs: http://www.cuhk.edu.hk
  - Numeric expressions: 12,345.67, 12’34”

Document Structure Analysis



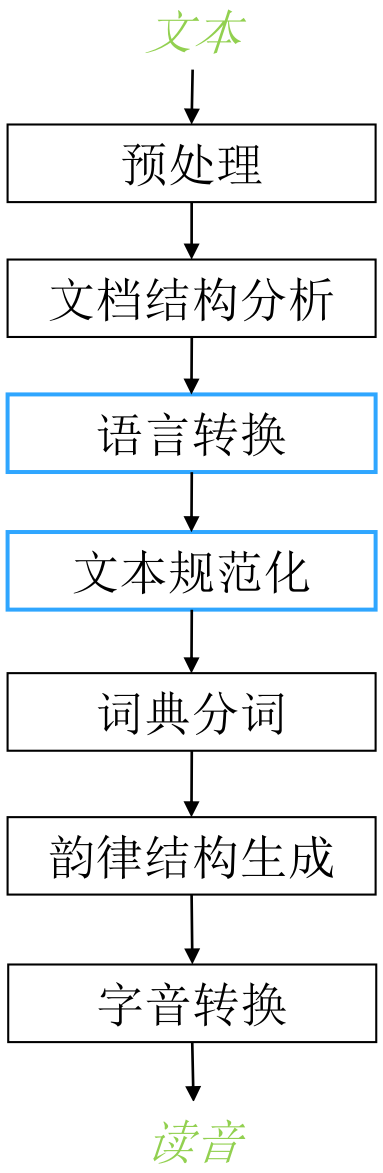
Module Input

```
<?xml version="1.0" encoding="UTF-16" ?>
<speak version="1.1" xml:lang="zh-cmn">
我说<w>道哥</w>，你还欠我HKD$10,000.00呢！
有冇搞錯！
</speak>
```

Module Output

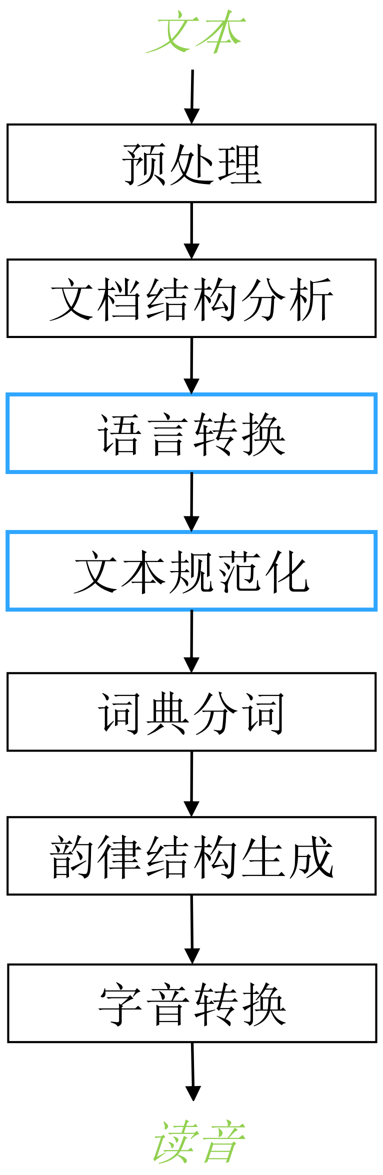
```
<speak version="1.1" ...... xml:lang="zh-cmn">
<p>
 <s xml:lang="zh-Hans">我说<w>道哥</w>，你还欠我
  <say-as interpret-as="measure"
  format="measure">HKD$10,000.00</say-as>呢！ </s>
</p>
<p>
 <s xml:lang="zh-Hant">有冇搞錯！ </s>
</p>
</speak>
```

Text Normalization



- **To convert all other written languages to engine-specific language**
  - 语言转换
    - 将其他书写语言转换为合成引擎相关的语言
- **To convert all written form of special construct into corresponding spoken form**
  - 文本规范化
    - Non-Standard Words
      - 逗点：数字中‘,’的处理
      - 句点：12.3、166.111.68.142
      - 量词：km、T等处理
      - 年份日期1998/07/20、97-10-10、1999.07.05等
      - 时间23:05:03、比值
      - 符号：-5℃、电话中BP机呼号、区号、转分机等、-5、80-100、减号等
      - 其他数字：公元、电话、电报等

Text Normalization



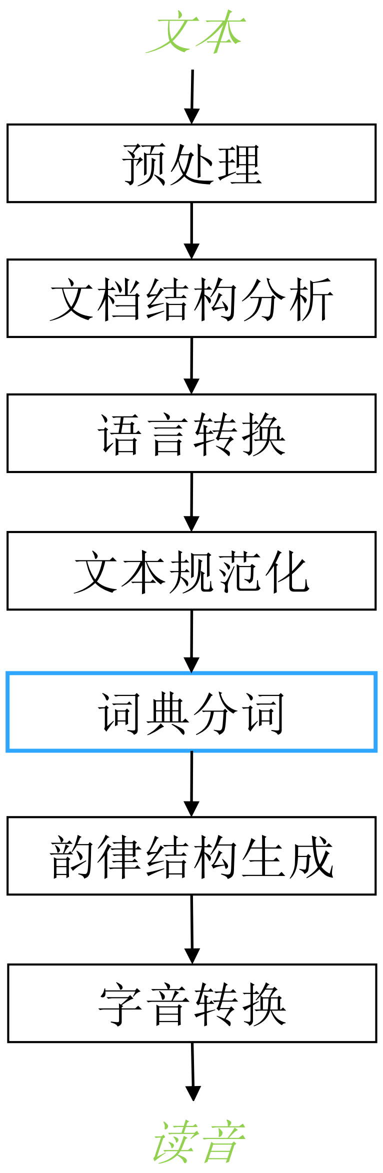
Module Input

```
<speak version="1.1" ...... xml:lang="zh-cmn">
  <p>
    <s xml:lang="zh-Hans">我说<w>道哥</w>，你还欠我
    <say-as interpret-as="measure"
      format="measure">HKD$10,000.00</say-as>呢！</s>
  </p><p>
    <s xml:lang="zh-Hant">有冇搞錯！</s>
  </p>
</speak>
```

Module Output

```
<speak version="1.1" ...... xml:lang="zh-cmn-Hans">
  <p>
    <s>我说<w>道哥</w>，
      你还欠我<w>港币</w><w>一万块</w>呢！</s>
  </p><p>
    <s>有没搞错！</s>
  </p>
</speak>
```

Word Tokenization



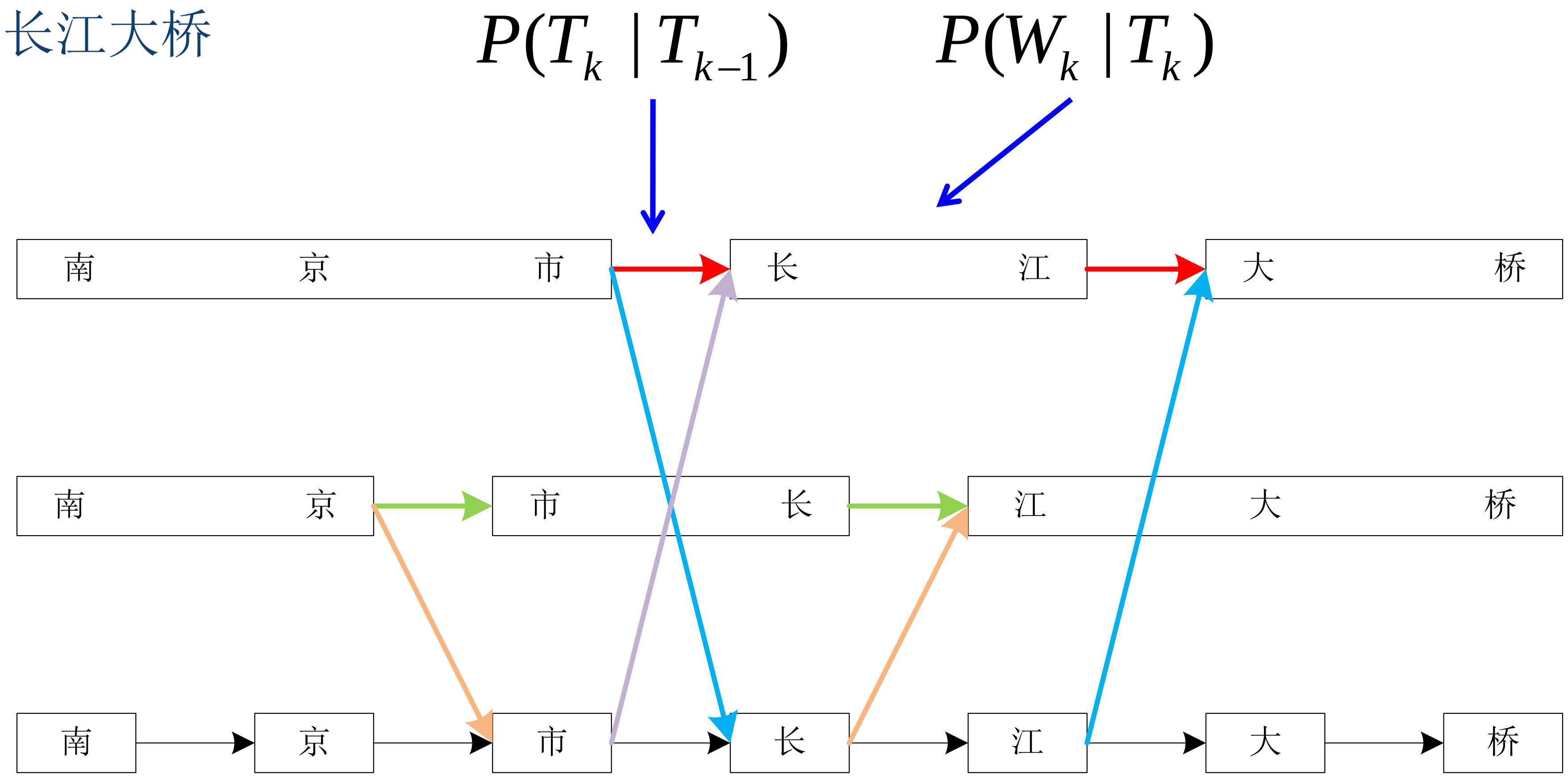
- **To tokenize sentence into words according to word tokenization lexicon**
  - 词典分词
  - 标记词性

POS Bi-gram based Word Tokenization Algorithm

- Same POS (part-of-speech, 词性) bi-gram model for both Putonghua and Cantonese TTS engines

POS Bi-gram based Word Tokenization

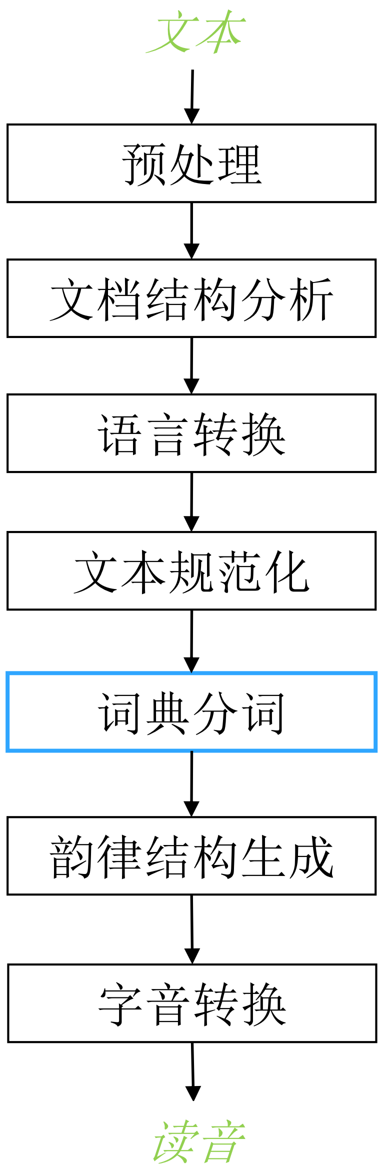
- **An example**
  - 南京市长江大桥



$P(W_k|T_k)$: the probability for the *k*th word being W_k given POS T_k.
$P(T_k|T_{k\text{-}1})$: the transition probability from previous POS $T_{k\text{-}1}$ to current POS T_k.

$$P(T_k \mid T_{k-1}) = \frac{F(T_k T_{k-1})}{F(T_{k-1})}, P(W_k \mid T_k) = \frac{F(W_k, T_k)}{F(T_k)}$$

Word Tokenization



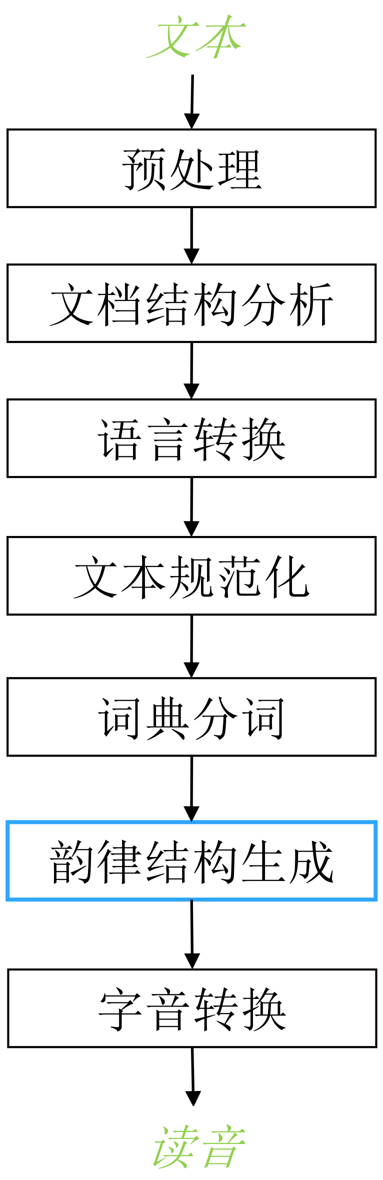
- **Module Input**

```
<speak version="1.1" ...... xml:lang="zh-cmn-Hans">
 <p>
  <s>我说<w>道哥</w>,
   你还欠我<w>港币</w><w>一万块</w>呢！ </s>
 </p><p>
  <s>有没搞错！ </s>
 </p>
</speak>
```

- **Module Output**

```
<s> ……
<w role="r" freq="5733">你</w>
<w role="d" freq="19396">还</w>
<w role="v" freq="360">欠</w>
<w role="r" freq="19875">我</w>
<w role="n" freq="98">港币</w>
<w role="m" freq="125">一万块</w>
<w role="y" freq="2092">呢</w>
<w role="wt" freq="1">！ </w>……
</s>
```

Prosodic Structure Generation



- 词典词（语法词）与韵律词不等同
  - 我 买 了 八 本 书。
  - 你 还 欠 我 港币 一万块 呢！

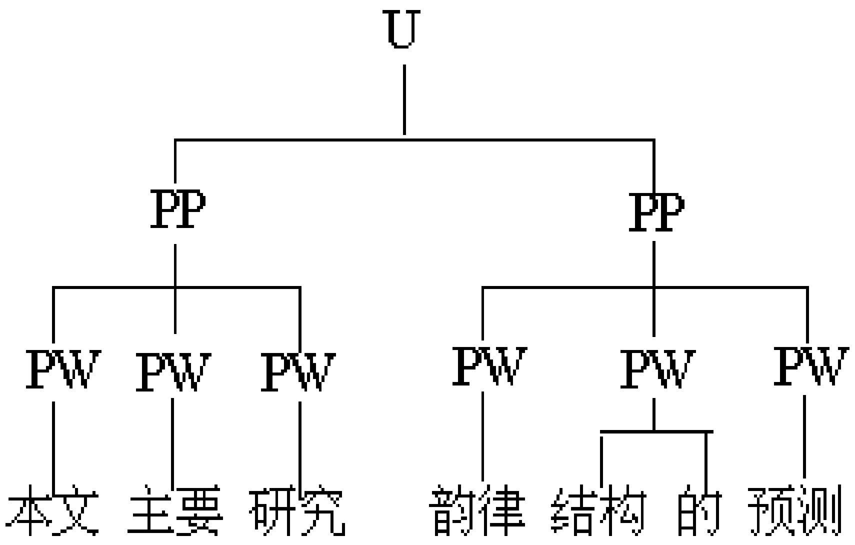
- **To generate prosodic structures (boundaries for prosodic word and phrase)**
  - 韵律结构生成

Prosodic Structure of Chinese

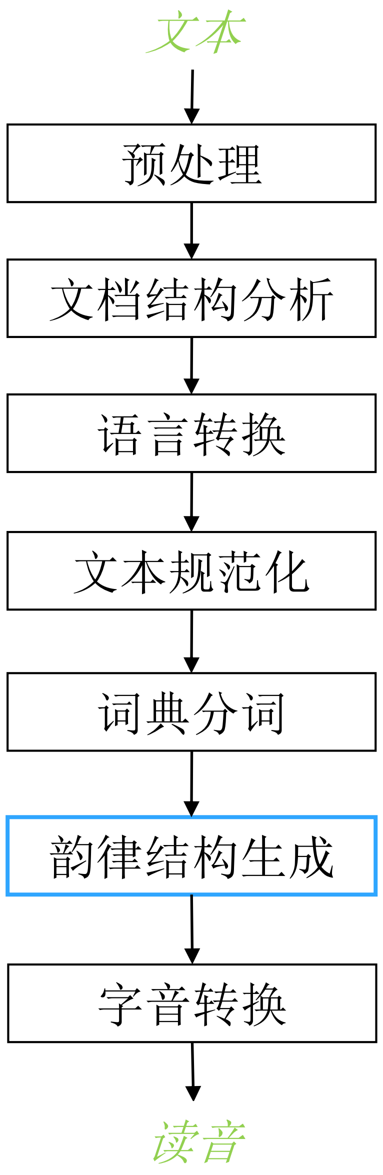
- **汉语韵律层级结构**
  - 语法词/词典词（lexicon word）
    - 从句法学的角度定义的基于词典得到的分词信息
  - 韵律词（prosodic word）
    - 一般为三个音节以下的语法词或词组，内部不出现节奏边界
  - 韵律短语（prosodic phrase）
    - 由一个或几个韵律词组成，具有相对稳定的短语语调模式和短语重音配置模式
  - 语调短语（intonation phrase）
    - 长于韵律短语。在语法上相当于较短的句子或较长的短语，语调短语之间有音高重设
  - 语句（utterance）
    - 以标点符号（句号、逗号、分号等）分割开的小句

Prosodic Structure of Chinese

- 汉语韵律层级结构
  - 语法词/词典词（lexicon word）
    - 从句法学的角度定义的基于词典得到的分词信息
  - 韵律词（prosodic word）
    - 一般为三个音节以下的语法词或词组，内部不出现节奏边界
  - 韵律短语（prosodic phrase）
    - 由一个或几个韵律词组成，具有相对稳定的短语语调模式和短语重音配置模式
  - 语句（utterance）
    - 以标点符号（句号、逗号、分号等）分割开的小句



Prosodic Structure Generation



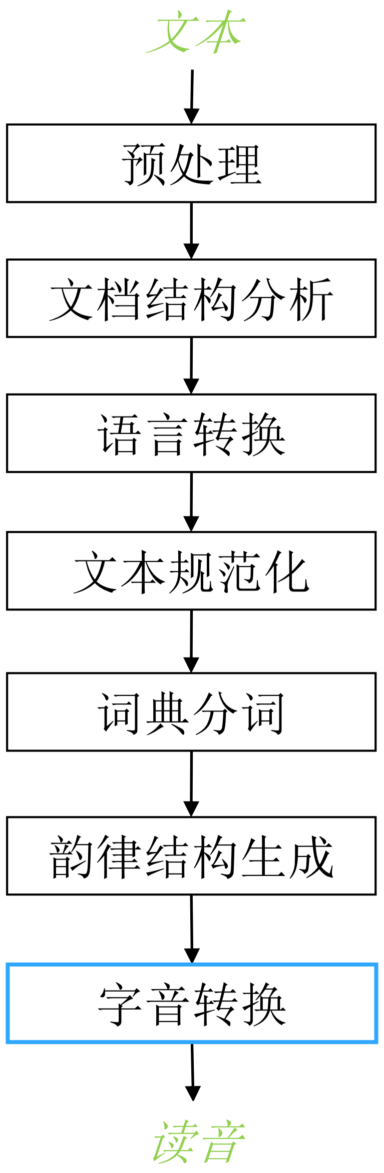
- **Module Input**

```
<s> ……
  <w role="r" freq="5733">你</w>
  <w role="d" freq="19396">还</w>
  <w role="v" freq="360">欠</w>
  <w role="r" freq="19875">我</w>
  <w role="n" freq="98">港币</w>
  <w role="m" freq="125">一万块</w>
  <w role="y" freq="2092">呢</w>
  <w role="wt" freq="1">！ </w>……
  </s>
```

- **Module Output**

```
<s> ……
  <break strength="strong" />
    <w role="r" freq="5733">你</w>
    <w role="d" freq="19396">还</w>
  <break strength="weak" />
    <w role="v" freq="360">欠</w>
    <w role="r" freq="19875">我</w>
  <break strength="medium" />
    <w role="n" freq="98">港币</w>
   <break strength="weak" />
     <w role="m" freq="125">一万块</w>
     <w role="y" freq="2092">呢</w>
  <break strength="strong" />
    <w role="wt" freq="1">！ </w>……
</s>
```

Grapheme-to-Phoneme Conversion



- **To derive the pronunciation for each word from pronunciation lexicon**
  - 字音转换
  - 1) 查发音字典
    - 从发音词典根据分词结果、词性获得对应读音
      - 种,zhong4,v,1210
      - 种,zhong3,q,8119
      - 种养,zhong4yang3,v,30
      - 种别,zhong3bie2,n,2
  - 2) 多音字处理
  - 3) 音变处理
      - “一、不”变调
        - 一个、不要
        - 第一、不准
      - 三声连读变调
        - 保卫、饱满、法宝、珠宝
        - 老虎， 555， 5599
      - 轻声
        - 爸爸、妈妈、姐姐、哥哥

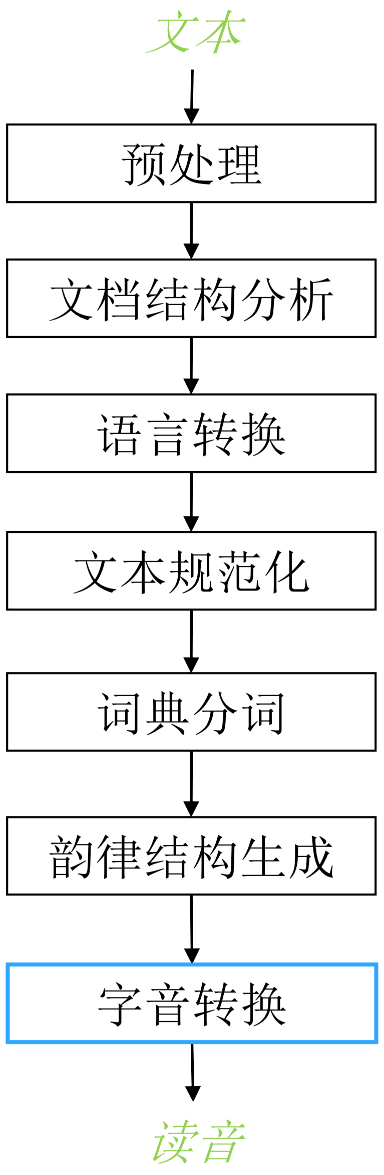
Homograph Disambiguation

- **Homograph (同形异义字) / Polyphone (多音字)**
  - Words with same spelling but different pronunciation
  - 形式上相同，但具有不同语法，语义功能的词。

- **多音字现象**
  - 250万字语料，多音字占8.95％
  - 举例如下：
    - 1. 我们种了茄子　动词，zhong4
    - 2. 一种新的算法　量词，zhong3
    - 3. 你真有种　名词，zhong3
    - 4. 各种各样　包含在词的内部，zhong3
    - 5. 种小明刚刚毕业　作为姓氏，chong2

- **多音字消歧**
  - 语义词典Hownet
  - 人工定义的上下文模式和语法信息
  - 问题：使用语义词典，需要语义标注；人工规则费时费力

Grapheme-to-Phoneme Conversion



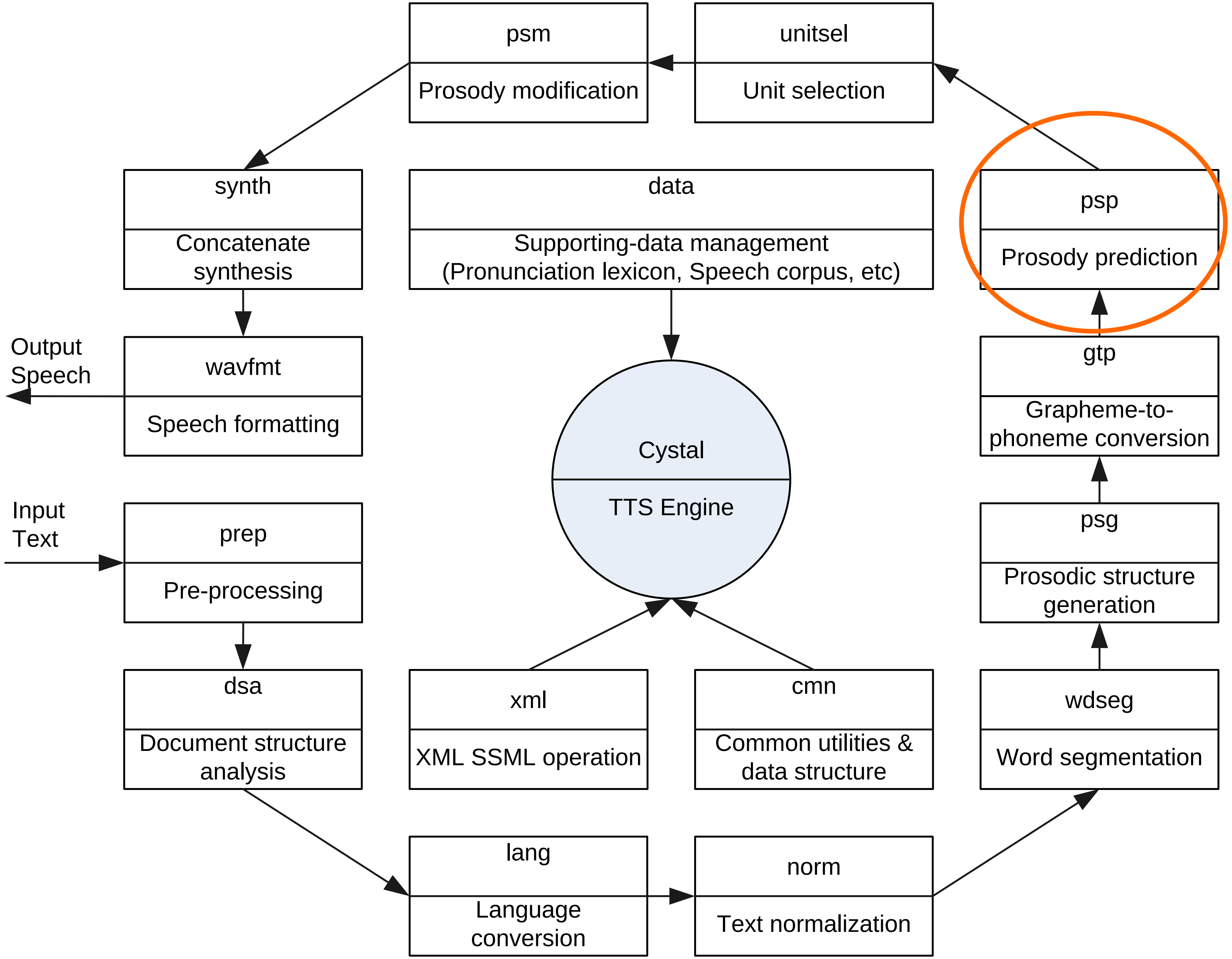
Module Output

```
<s> ……
  <break strength="strong" />
    <w role=“r”><phoneme alphabet=“pinyin” ph=“ni3”>你
      </phoneme></w>
    <w role="d"><phoneme alphabet="pinyin" ph="hai2">还
      </phoneme></w>
  <break strength="weak" />
    <w role="v"><phoneme alphabet="pinyin" ph="qian4">欠
      </phoneme></w>
    <w role="r"><phoneme alphabet="pinyin" ph="wo3">我
      </phoneme></w>
  <break strength="medium" />
    <w role="n">
      <phoneme alphabet="pinyin" ph="gang3 bi4">港币
      </phoneme></w>
  <break strength="weak" />
    <w role="m“>
      <phoneme alphabet=“pinyin” ph=“yi2 wan4 kuai4”>一万块
      </phoneme></w>
    <w role="y"><phoneme alphabet="pinyin" ph="ne0">呢
      </phoneme></w>
    <w role="wt">！ </w>……
</s>
```

韵律预测

Prosody Prediction

Crystal: An Example

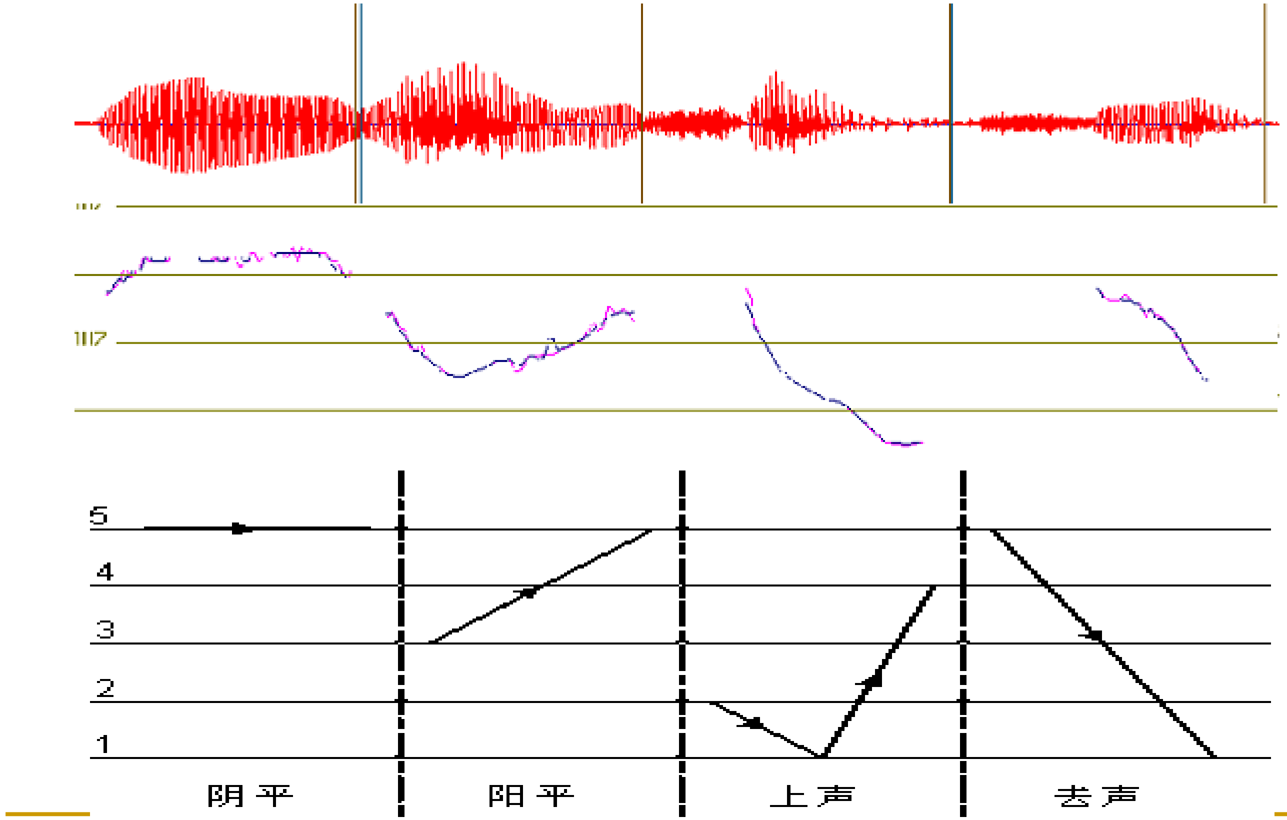


Prosodic Features and Prediction

- **Prosodic Phrasing: 韵律节奏**
  - Need to break utterances into phrases
  - Punctuation is useful, not sufficient
  - Prosodic structure generation: 韵律结构预测
- **Pitch Contour: 基频曲线**
  - Predictions of accents
    - which syllables should be accented / stressed (重读)
  - Realization of F0 contour
    - given accents/tones, generate F0 contour
- **Duration: 时长**
  - Predicting duration of each speech unit
    (e.g., phoneme for English, syllable for Chinese, etc.)
- **Energy: 能量**
  - Related to loudness

Pitch / Tone: 音高/声调

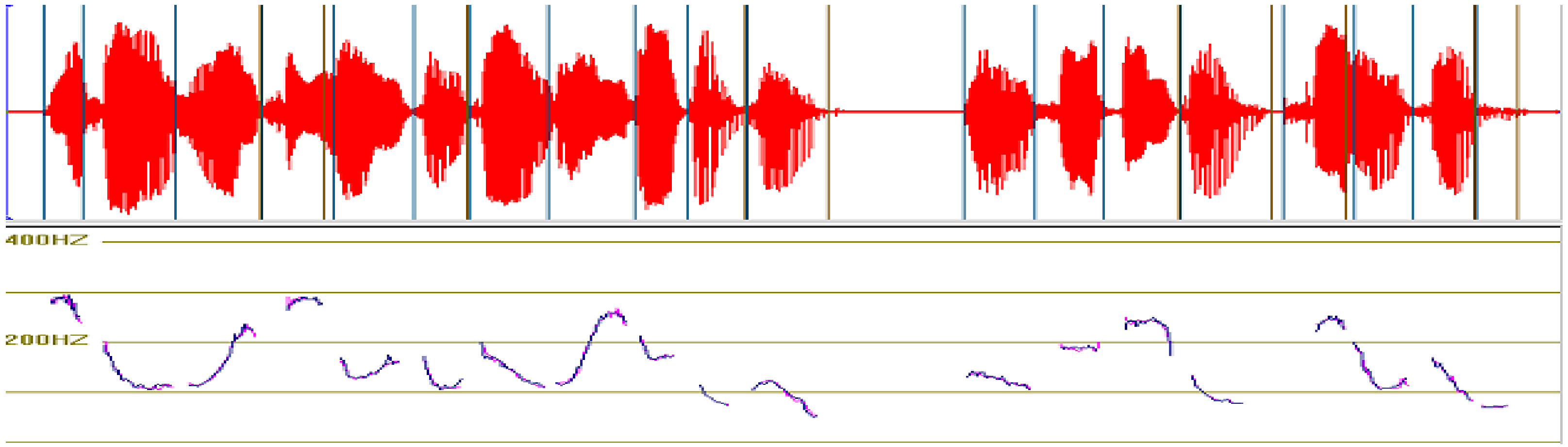
- **声调 (Tone) 是音节音高频率的包络 (Pitch contour)**
- **声调的区别特征：**
  - 高、低、抬高、上升、下降、下凹
  - 根据调型的差异，汉语普通话分为阴(平)、阳(平)、上(声)、去(声)四个调类，另有轻声调

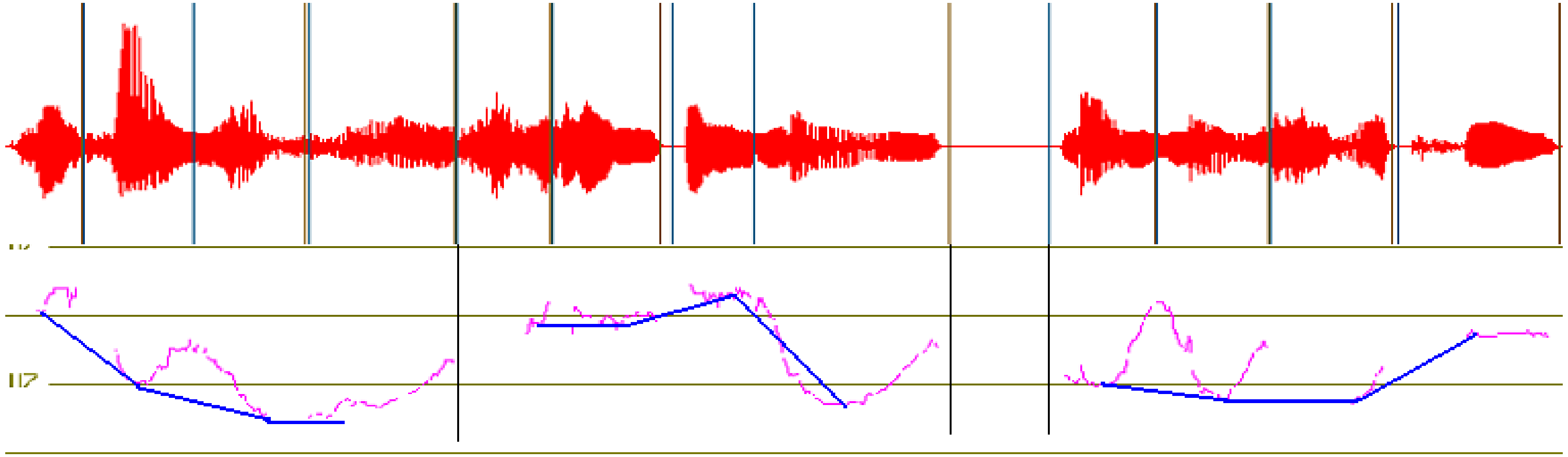


Intonation：语调

- **语调**
  - 陈述、疑问、祈求、商量、感叹、命令等
- **语调与声调的共存是同种超音段成分（音高）的共时结合**
  - 声调类型是词的结构的一部分，是对音高的调节
  - 而语调是对音高的再调节
  - 语调短语
    - 对于长句来说，一般可分为数个短语
    - 短语自成单元，具有完整的规则调群
  - 语调的变化
    - 语调的变化伴随着句子的节奏、速度的改变
    - 其声学特征表现为音高、音域、时长、音强和停顿的变化

汉语的语调



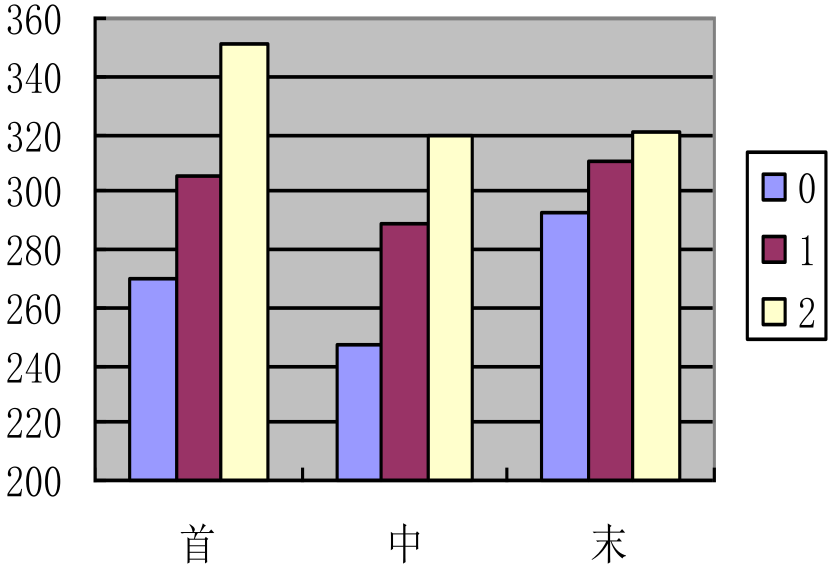


韵律分析：Prosody Analysis

- 韵律分析
  - 基于语音数据的语音研究，就是对语料库大量的语音数据进行基本声学参数的统计
  - 根据结果挖掘发现语音的规律和知识，测试和检验语音理论
- 基本声学参数
  - 时长
  - 基频

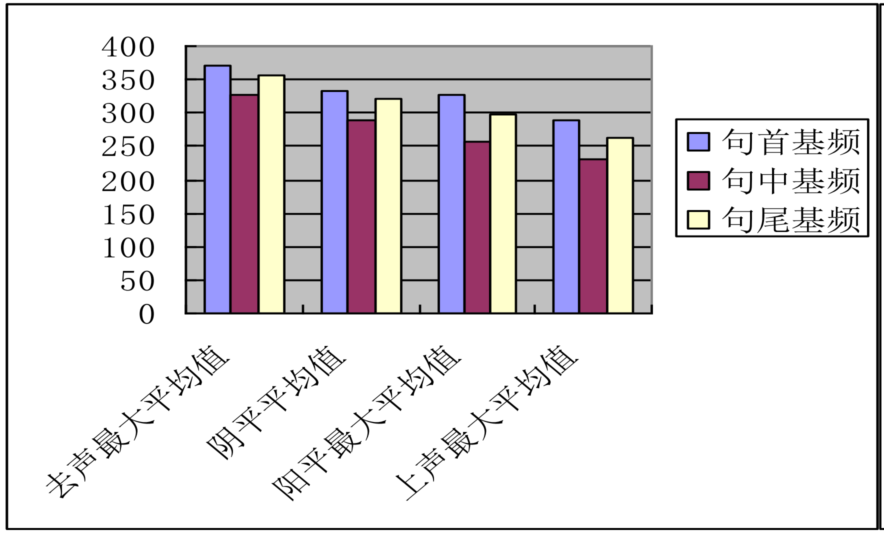
影响音节时长的因素

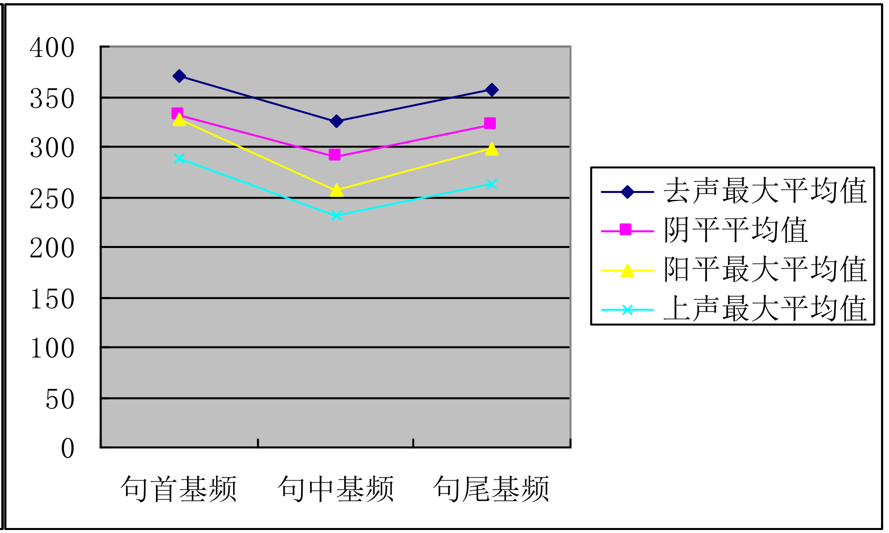
- 在连续语流中，影响音节的时长的因素：
  - 发音速度
  - 音节在句中的位置
  - 音节在韵律结构中的地位
  - 该音节是否是重读音节



	音节	韵律词边界	韵律短语边界	语调短语边界	语句边界	平均
语料库1	273 ms	325 ms	392 ms	388 ms		
语料库2	215.5 ms	235ms	297.1 ms	314.6 ms		
语料库3	221 ms	210 ms	298 ms	325 ms	307ms	245ms

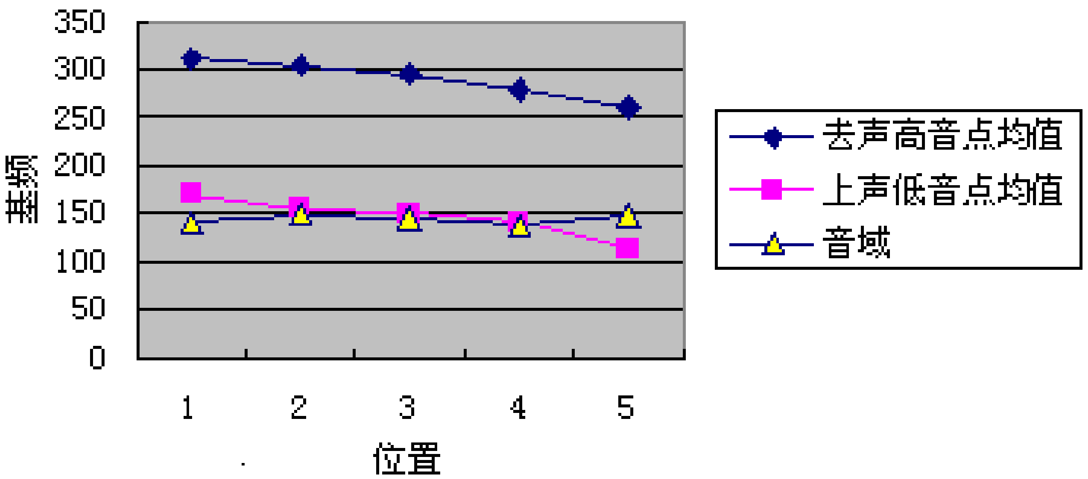
基频与声调、句中位置的关系





	句首基频	句中基频	句尾基频	平均基频
去声最大平均值	**371**	**326**	**356**	335
阴平平均值	**332**	**290**	**322**	315
阳平最大平均值	**327**	**257**	**298**	271
上声最大平均值	**288**	**231**	**263**	242
阳平最小平均值	220	156	199	169
上声最小平均值	182	97	145	124
去声最小平均值	229	145	213	164

基频、音域与短语位置的关系



	音节位置	短语首音节	短语次首音节	短语中音节	短语次尾音节	短语尾音节
阴平	基频均值	300 Hz	289 Hz	282 Hz	252 Hz	240 Hz
阳平	高音点均值	273 Hz	264 Hz	255 Hz	232 Hz	230 Hz
	低音点均值	199 Hz	192 Hz	186 Hz	172 Hz	162 Hz
上声	高音点均值	225 Hz	226 Hz	209 Hz	198 Hz	185 Hz
	低音点均值	**170 Hz**	**155 Hz**	**150 Hz**	**141 Hz**	**113 Hz**
去声	**高音点均值**	**311 Hz**	**304 Hz**	**295 Hz**	**278 Hz**	**261 Hz**
	低音点均值	209 Hz	192 Hz	192 Hz	175 Hz	138 Hz

影响声学特征的因素：语言学上下文信息

- **音节幅度**
  - 音节在语句中的位置
- **音节时长**
  - 声韵母、声调、音节在语句中的位置
- **音节基频（声调）**
  - 声调、相邻音节的声调
    - 相邻前音节的韵母类型和声调类型
    - 相邻后音节的声母类型和声调类型
  - 音节在语句、短语中的位置
    - 音节所在韵律短语信息（音节数、在句中位置、重音类型、距前一个重音距离和距后一个重音距离）

- **当前音节信息**
  - 声母类型、韵母类型、声调类型、在词中位置、与前音节耦合度和与后音节耦合度
- **语句信息**
  - 语句类型和韵律短语个数

韵律预测：Prosody Prediction

- **韵律预测**
  - 根据韵律分析的结果，对声学特征参数（基频、时长、能量）与上下文信息之间的关系进行统计建模
  - 合成时，根据待合成的上下文信息，通过上述模型预测得到相应的声学特征参数
- **基本声学参数**
  - 时长
  - 基频

Prosody Prediction

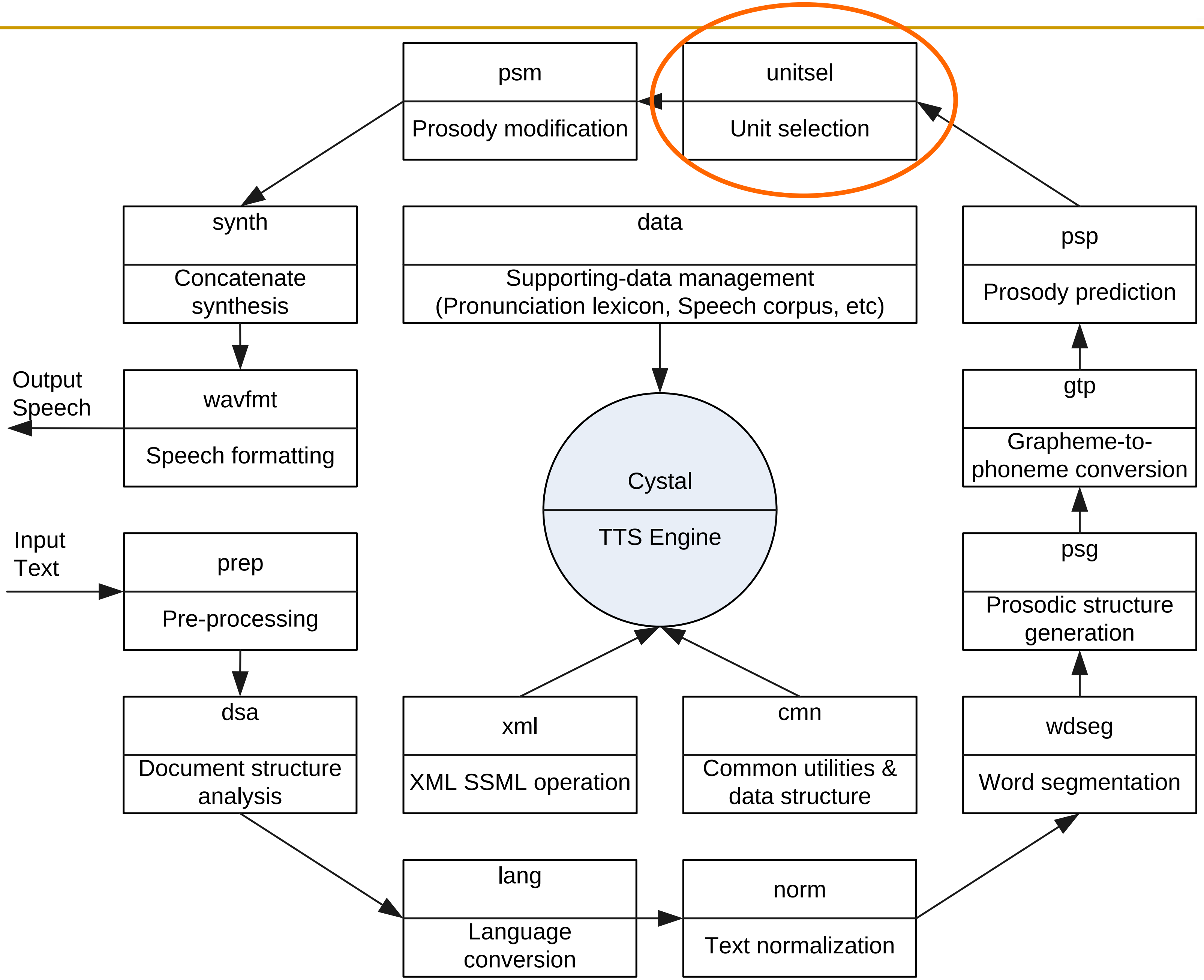
- **Module Output**

```
<s> ……
  <break strength="strong" />
    <w role=“r”>
      <unit>
        <prosody duration="262ms">
          <phoneme alphabet="pinyin" ph="ni3">你</phoneme>
        </prosody>
      </unit>
    </w>
    <w role="d">
      <unit>
        <prosody duration="299ms">
          <phoneme alphabet="pinyin" ph="hai2">还</phoneme>
        </prosody>
      </unit>
    </w>
  <break strength="weak" />
    ……
    <w role="wt">！ </w>……
</s>
```

语音基元选取

UNIT SELECTION

Crystal: An Example



Unit Selection

- 基元选取
  - 根据文本分析和韵律预测的结果从语音数据库中选取合适的语音基元
- 基元选取考虑的参数
  - 当前音节的读音
  - 前后音节的读音（前音节韵母、后音节声母）
  - 前后音节的声调
  - 韵律层级结构信息（韵律词、韵律短语、语句等）
  - 目标韵律特征参数（基频、时长、重读等）

Speech Library：语音数据库

- **单样本 V.S. 多样本**
  - 样本的听感差异

我抓你。你唱歌。 纷纷扬扬的

059，210，593，716，937，482，105，371，164，821，648，

- **引起样本听感差异的因素（韵律参数集）**
  - 位置
  - 相邻声调
  - 相邻声韵母
  - ……

- **如何得到更高语音自然度的合成语音**
  - 选择不同韵律特征的多个语音基元
  - 选取拼接（符合韵律特征的）最优基元

Unit Selection: 语音基元选取

- **基于规则（少量样板）**
  - 位置
  - 相邻声调
  - 相邻声韵母

我抓你。你唱歌。　　纷纷扬扬的

059，210，593，716，937，482，105，371，164，821，648，

- **基于大规模语音数据**
  - 选择不同韵律特征的多个语音基元
  - 语音数据库大小 vs. 合成语音质量
  - 选取拼接最优基元

语音基元的选取方法

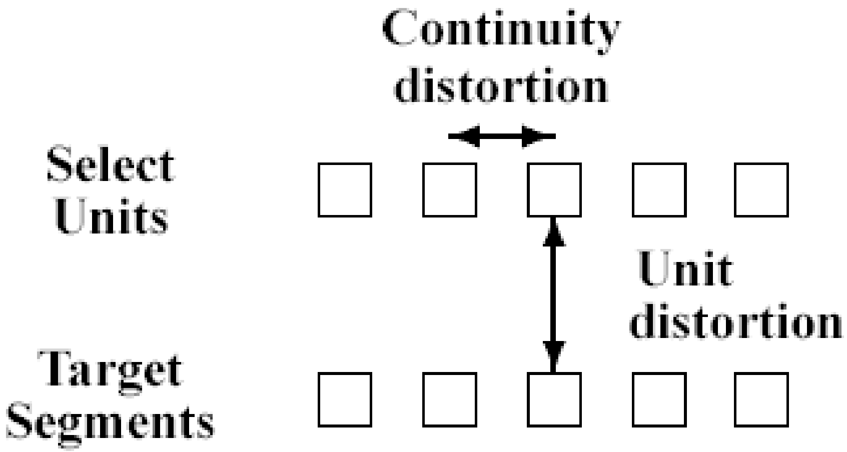
- **基于韵律代价函数的语音基元选取**
  - 韵律代价函数：匹配代价，拼接代价
  - 基于Viterbi算法的最优基元路径搜索
  - 基元选取权重的设定

- **基于决策树的语音基元选取**
  - 决策树的定义
  - 决策树的训练：问题集
  - 决策树的训练：距离函数
  - 基于决策树的语音基元选取

Target Cost and Joint Cost

- **Target Cost: 匹配代价**
  - 描述待合成音节的目标参数和语音库中每个具有特定语音和韵律上下文的语音候选单元之间的不匹配程度

$$C^t(t_i,u_i)=\sum_{j=1}^{p} w_j^t C_j^t(t_i,u_i)$$

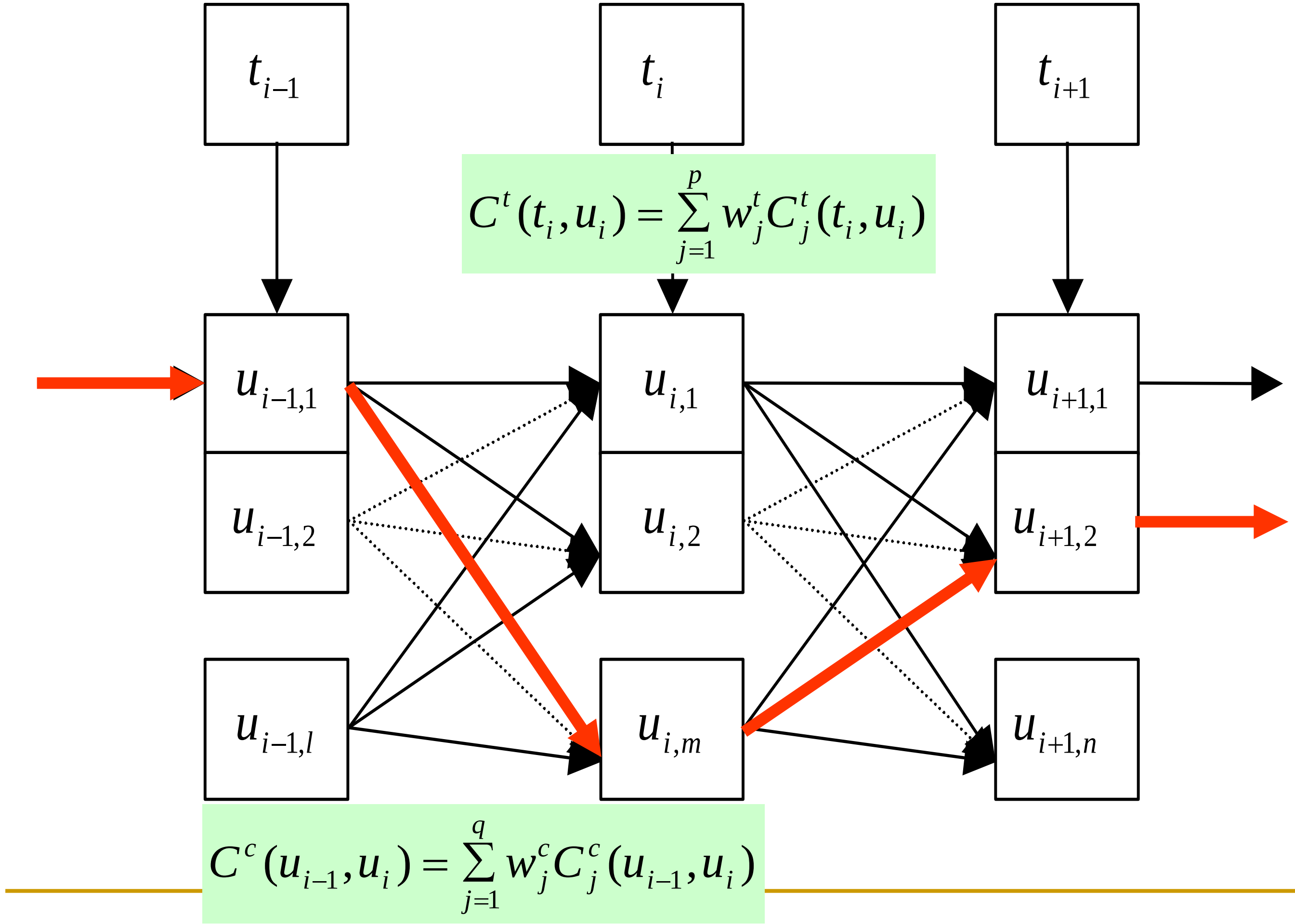


- **Joint Cost: 拼接代价**
  - 描述相邻的语音候选单元在拼接时它们之间的连接和匹配情况

$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1},u_i)$$

Viterbi Algorithm for Unit Selection

- **基元选取过程**
  - 根据匹配代价得到一系列的候选基元以后，然后利用Viterbi搜索算法，从所有路径中搜索一条最佳的（韵律代价最小的）路径，这条路径上的单元即作为最终用于合成拼接的单元



$$C^t(t_i,u_i)=\sum_{j=1}^{p} w_j^t C_j^t(t_i,u_i)$$

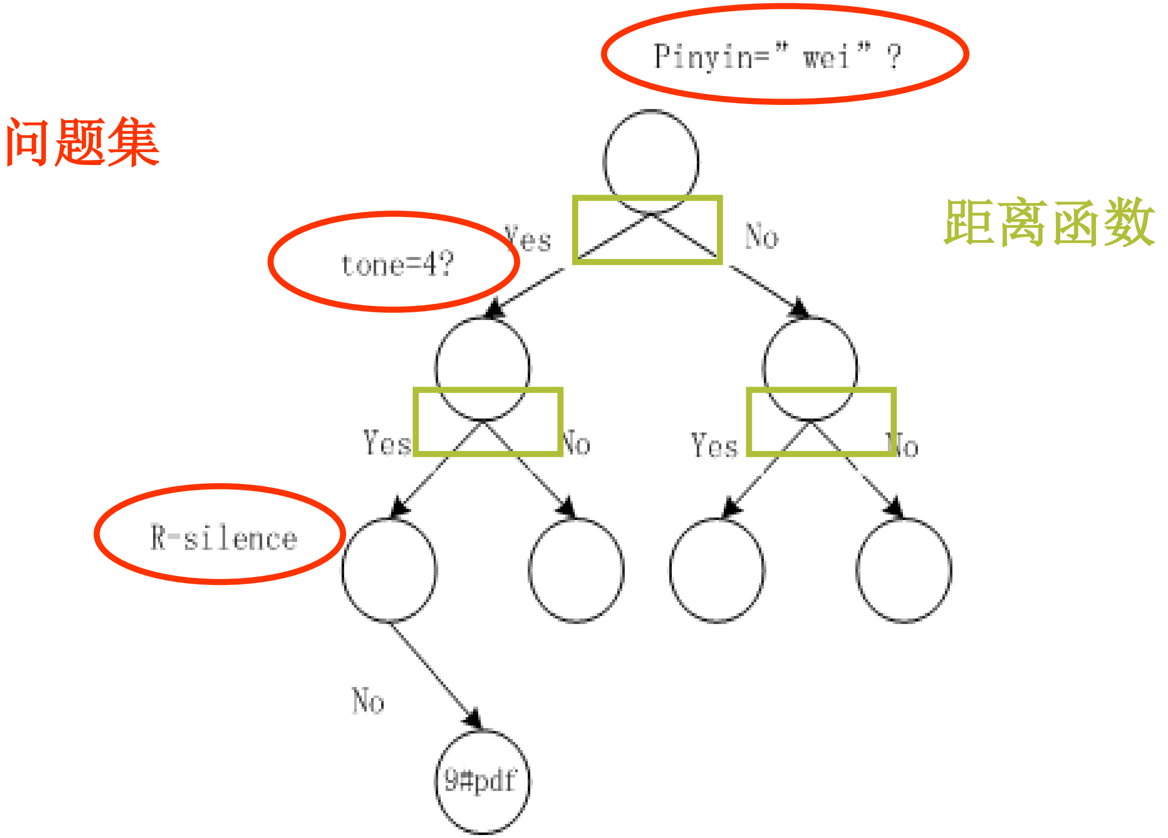
$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1},u_i)$$

CART based Unit Selection

- 基元选取
  Unit Selection
  - Query the CART with the phonetic context and prosodic context which are produced by the text analysis module to find the proper leaf node.

CART

- 分类与回归树 CART
  - Classification and Regression Tree
  - CART树的训练（构造）：节点分裂准则



问题集

距离函数

CART树的训练 – 问题集

- **语境上下文信息: Context Information**

- **当前音节在所在韵律词中的位置PosInWord**
  - 取head、body、tail三个值；
- **当前音节在所在韵律短语中的位置PosInPhrase**
  - 取head、body、tail三个值；
- **前音节的声调类型PreTone**
  - 取high、low、neutral、null四个值；
- **后音节的声调类型PostTone**
  - 取high、low、neutral、null四个值；
- **前音节的韵母PreFinal**
  - 前面音节的韵母；
- **后音节的声母PostInitial**
  - 后面音节的声母。

CART树的训练 – 距离函数

- 分裂准则
  - 节点分裂好坏的判断准则
  - 计算声学特征参数的距离

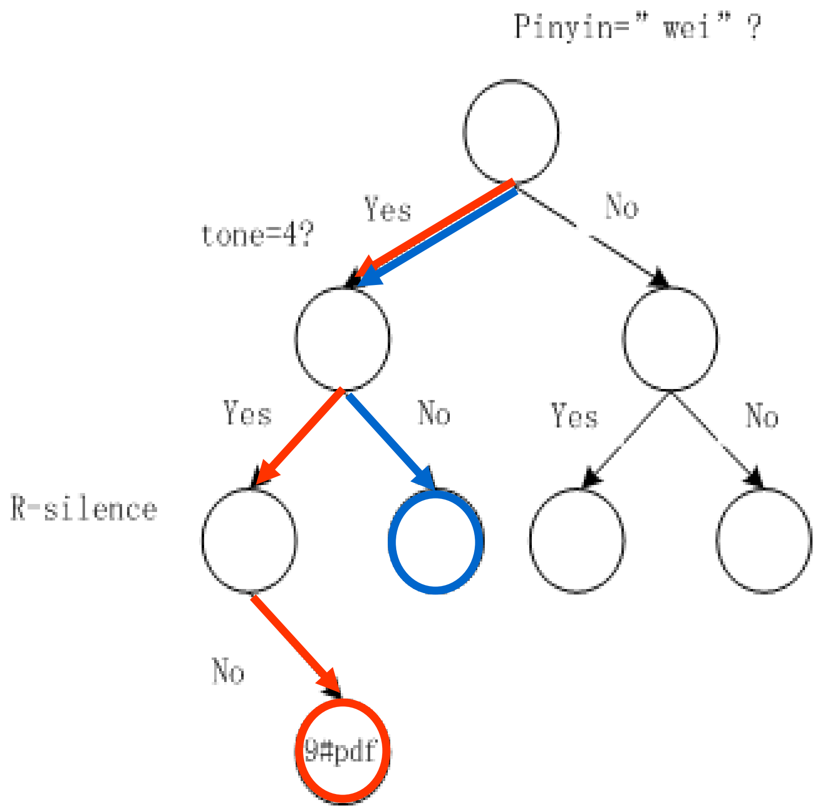
- 时长预测树
  - 音节的时长 D
- 能量预测树
  - 音节的均方根能量 E
- 基频曲线预测树
  - 音节的基频向量 $\mathbf{P} = \{p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9\}$

- 这里声学特征参数与音库裁剪算法中一致，用来计算两个样本之间的距离。

CART based Unit Selection

- 为临帖他还远游
  wei4 lin2 tie4 …
  - Pinyin = “wei”
  - Tone = 4
  - R-silence = False

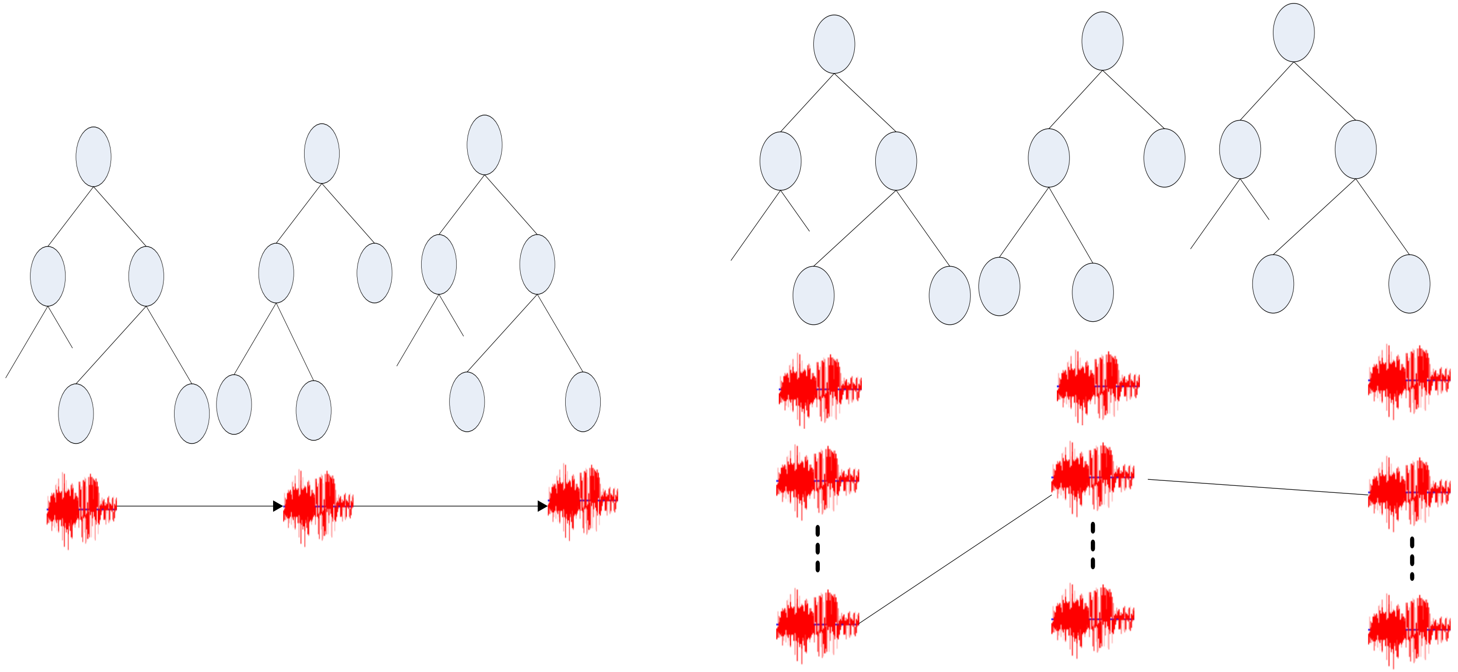
- 伟大的祖国
  wei3 da4 de0…
  - Pinyin = “wei”
  - Tone = 3



Multiple Instances in Leaf Node

- **单样本 (Single instance):**
  - 每个叶子节点仅保留一个样本

- **多样本 (Multiple instances):**
  - 每个叶子节点保留多个样本，可进一步基于代价函数及动态规划算法找到最佳样本序列
  - 需要更多存储空间及计算复杂度



Unit selection

- **Module Output**

```
<s> ……
  <break strength="strong" />
    <w role="r">
      <unit uid="ni3:0">
        <prosody duration="262ms">
          <phoneme alphabet="pinyin" ph="ni3">你</phoneme>
        </prosody>
      </unit>
    </w>
    <w role="d">
      <unit uid="hai2:2">
        <prosody duration="299ms">
          <phoneme alphabet="pinyin" ph="hai2">还</phoneme>
        </prosody>
      </unit>
    </w>
  <break strength="weak" />
    ……
    <w role="wt">！ </w>……
</s>
```

语音合成

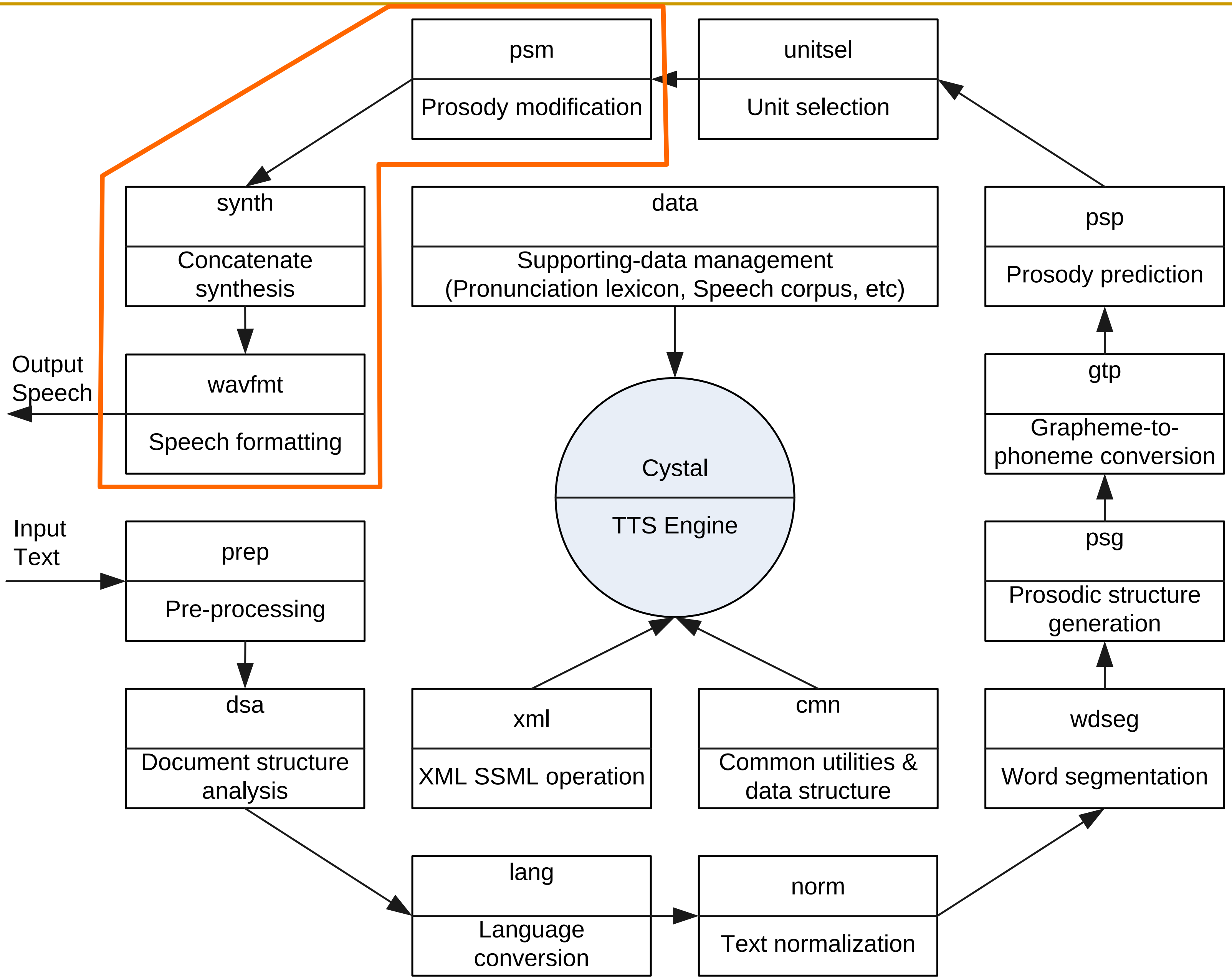
1) 韵律修改

2) 拼接合成

3) 格式转换输出

Speech Synthesis

Crystal: An Example



Speech Concatenate Synthesis

- **波形拼接语音合成**
  - 选取适当的合成基本单元（语音基元），拼接起来，得到整体的合成语音，播放。

语句、短语、词、音节、声母-韵母、音素

如何进行拼接合成以满足特定的韵律特征要求：基频高低、时长快慢、能量高低

他五点钟值班

Speech Concatenate Synthesis

- **波形拼接合成需要解决的主要问题**
  - 语音基元的选取
    - The prosodic information of the selected speech units does NOT necessarily match the target requirement.
    - The boundary of the adjacent units should be smoothed.
  - 修改基元的时域特征
    Prosody Modification
    - Pitch modification
    - Duration modification
    - Energy modification
  - 修改基元的谱特性
    - Unit edges: join smoothly and pitch-synchronously
  - 将不同的基元拼接以产生合成语音

Prosody Modification

- **Energy Modification**
  - Easy to implement: modify amplitude directly
- **Duration Modification**
  - Duplicate / remove parts of the signal
- **Pitch Modification**
  - Resample to change pitch
- **Requirement of Prosody Modification**
  - Modify pitch and duration independently
  - While changing sample rate modifies both

Speech Formatting

- **波形数据格式转换模块**
  - 将拼接合成得到的原始波形数据进行格式转换处理，比如

  - 格式转换：改变采样率、改变量化精度等
  - 生成wave文件：加上wave文件头
  - 生成mp3文件：进行压缩
  - 播放wave数据：直接通过声卡播放
  - 等

References

- *Progress in Speech Synthesis*, Jan P.H. van santen, et al., Springer-Verlag New York, Inc., 机械工业出版社(影印版)
- *语音合成 (中译本)* , 蔡莲红等, 机械工业出版社, 2003
- *An Introduction to Text-to-Speech Synthesis*, Thierry Dutoit, Kluwer Academic Publishers, The Netherlands, 1997
- *Multilingual Text-to-Speech Synthesis: the Bell Labs Approach*, Richard Sproat [edt.], Kulwer Academic Publishers, The Netherland, 1998
- *Text to Speech Synthesis: New Paradigms and Advances*, Narayanan Shrikanth, Alwan Abeer, Prentice Hall, USA, 2004
- SSML: Speech Synthesis Markup Language, [online] http://www.w3.org/TR/speech-synthesis/