可视语音合成

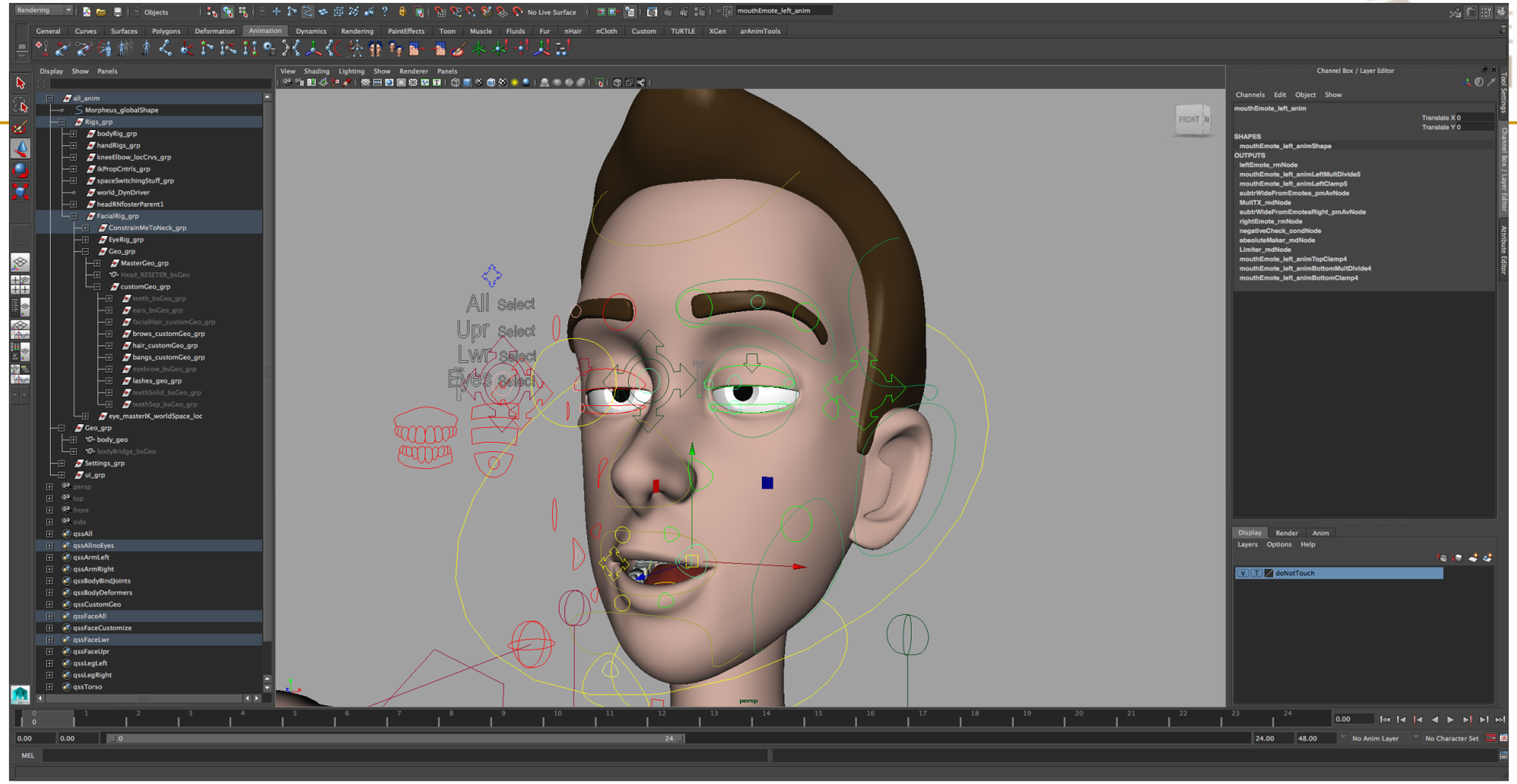
Visual Speech Synthesis

清华大学深圳研究生院

吴志勇

zywu@sz.tsinghua.edu.cn

MEDIA JOINT Research Center



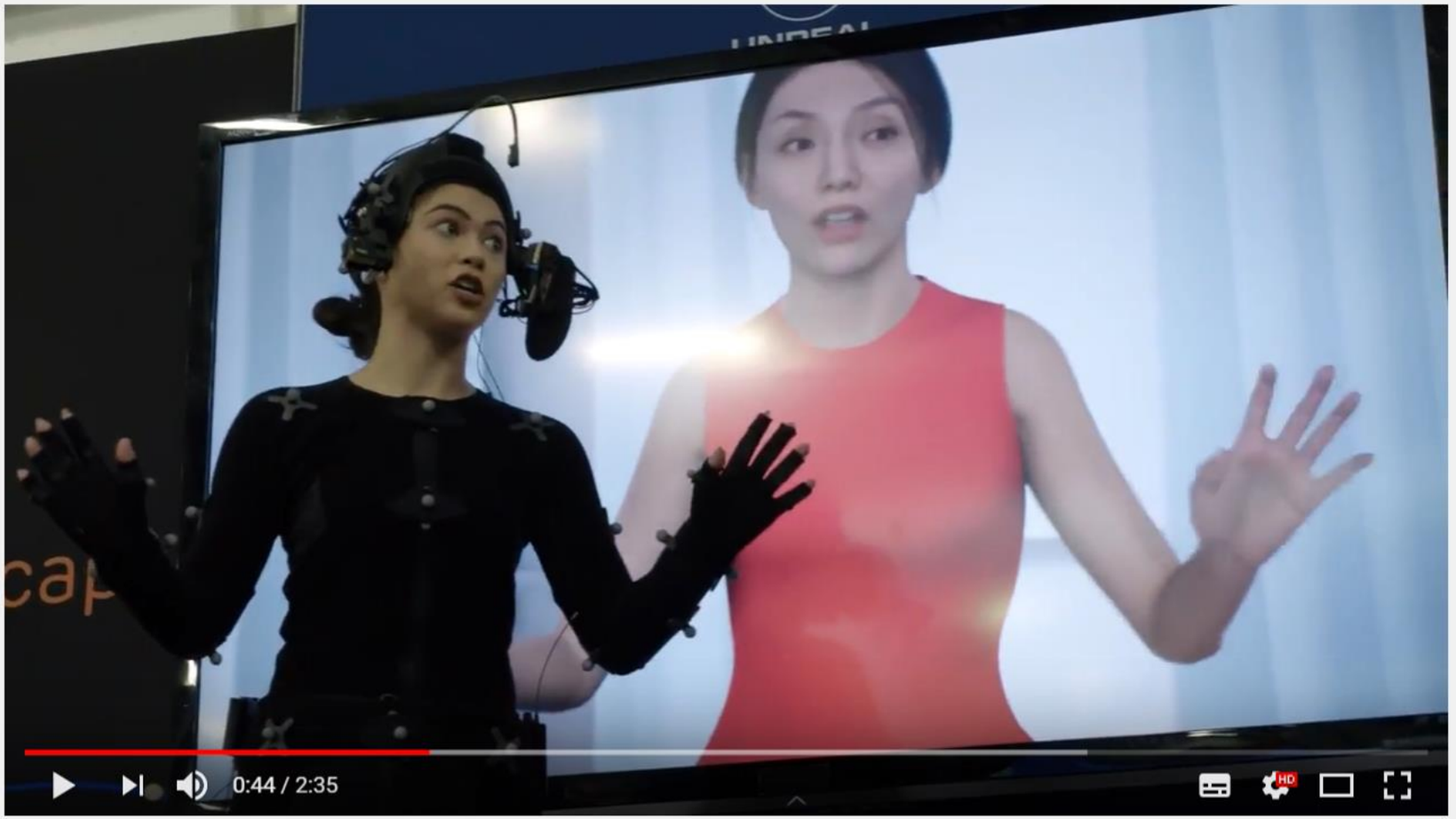
- Animation artists spend ≥50% time on face
  - Mostly eyes & mouth
  - Very tedious

We'll focus on mouth & speech.

Visual Effect in Movies

Visual Effect in Movies

From Movie to Avatar



Cubic Motion presents "Siren" at GDC 2018

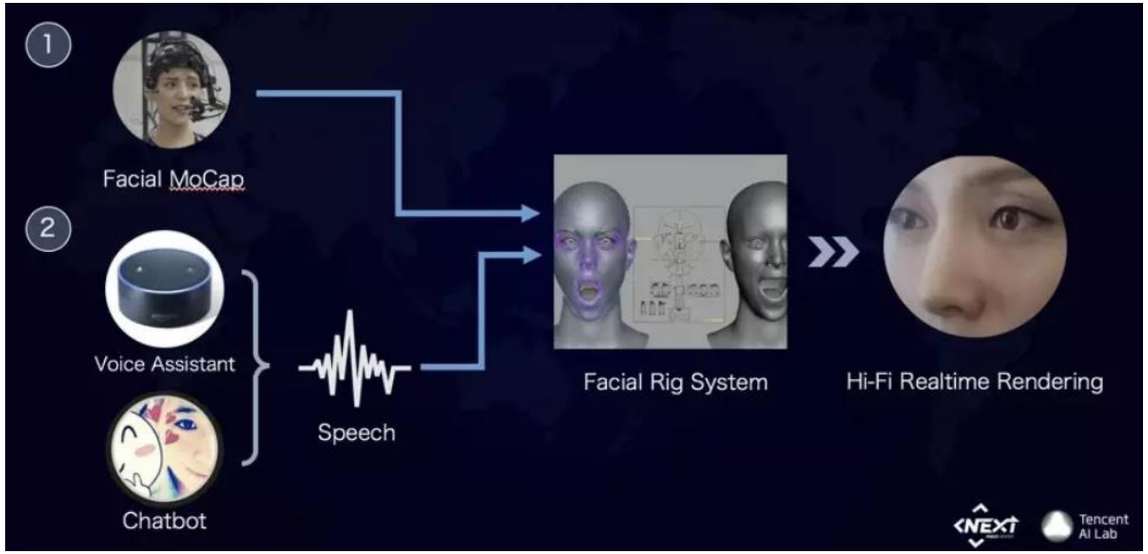
https://www.youtube.com/watch?v=mIFiftCLQsc

From Movie to Avatar

https://www.youtube.com/watch?v=NW6mYurjYZ0

Siren AI at SIGGRAPH Asia 2018

What's Behind Siren AI?



- **基于虚拟人的语音交互是一个复杂的过程：语音激活检测(VAD)，语音识别(ASR)，自然语言处理(NLP)，语音合成(TTS)，语音驱动面部动画(ADFA)**
  - 在虚拟人Siren的研究中，基于腾讯AI解决方案，主要攻坚语音驱动面部动画（ADFA）相关技术，以解决实时驱动虚拟人的技术难点。
  - Siren具有精巧的Rig Logic，能将80维左右的RigControls参数映射到数千维的脸部Rig Elements参数（Blend Shapes, Joints, Wrinkle Maps, etc.）。
  - 基于规则的映射，将语音驱动模型输出并抽象到80维左右，大大缩减模型规模，降低训练难度。训练时，在面部动捕的同时采集音视频数据和Rig Controls序列数据，离线处理成一一对应的训练数据，采用TimeCode和专业的音视频采集设备数据解决对齐和掉帧的问题。

可视语音合成的分类

- 可视语音合成（Visual Speech Synthesis）
- 虚拟说话人（Talking Avatar）

- 从动画生成的角度
  - 基于参数控制的可视语音合成
  - 基于数据驱动的可视语音合成

- 从输入控制的角度
  - 文本驱动的可视语音合成（ Text-To-Visual-Speech, TTVS ）
  - 语音驱动的可视语音合成（ Audio Driven Facial Animation, ADFA）

可视语音合成关注什么

- 口型动作（Lip Movement）
- 脸部表情（Facial Expression）
- 头部动作（Head Movement）
- 肢体动作（Gestures）

以口型动作为例……

主要内容

- **视位（Viseme）的概念**
- **基于参数控制的可视语音合成**
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- **基于数据驱动的可视语音合成**
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- **基于深度学习技术的可视语音合成**

主要内容

- **视位（Viseme）的概念**
- **基于参数控制的可视语音合成**
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- **基于数据驱动的可视语音合成**
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- **基于深度学习技术的可视语音合成**

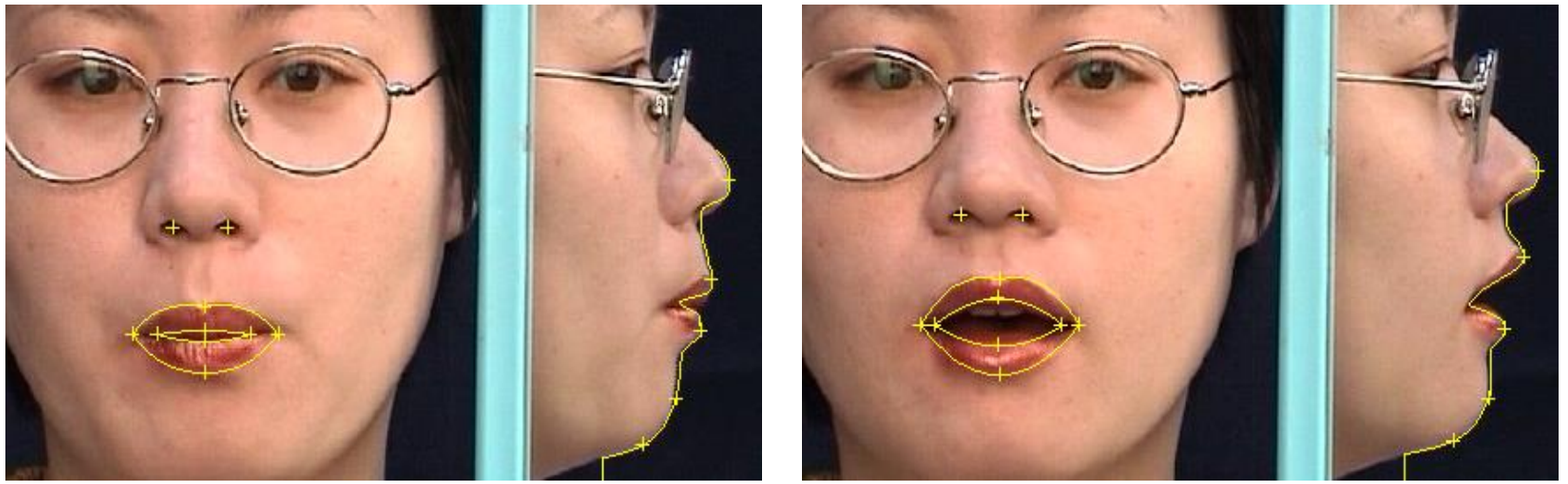
Viseme: 视位/视素

- **Viseme: Visual phoneme**
  - A **viseme** is a representational unit used to classify speech sounds in the visual domain, corresponding to the phoneme in the aural domain.
  - 语音在视觉域的最小单元，同听觉域的音素相关
  - A **viseme** describes the particular facial and oral positions and movements that occur alongside the voicing of phonemes.
  - 描述了在发某个特定音素时对应的可视发音器官的形状、位置、动作。

Chen, T. (2001). *Audiovisual speech processing.* IEEE Signal Processing Magazine, 9–31.

视位与可视语音合成

- 视位（Viseme = Visual + Phoneme）

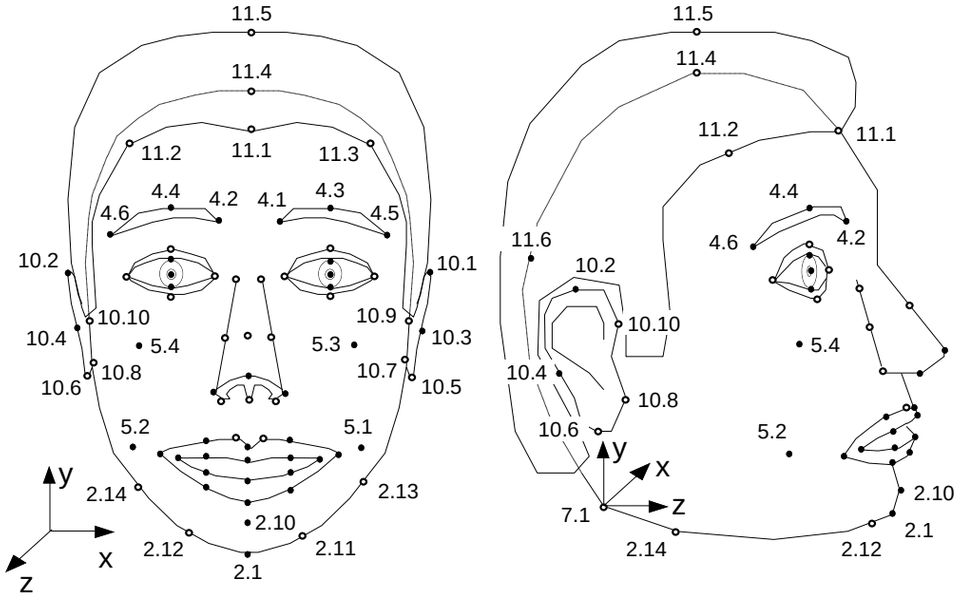


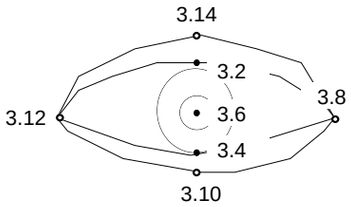
自动提取的静态视位/b/和/a/

主要内容

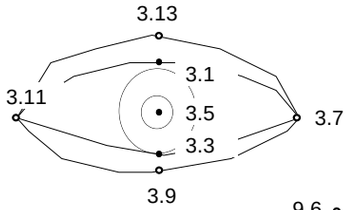
- 视位（Viseme）的概念
- **基于参数控制的可视语音合成**
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- 基于数据驱动的可视语音合成
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- 基于深度学习技术的可视语音合成

MPEG-4 Facial Definition Points (FDP)

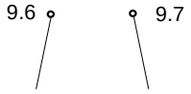




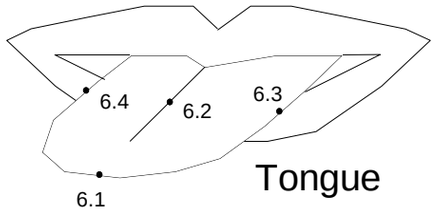
Right eye



Left eye



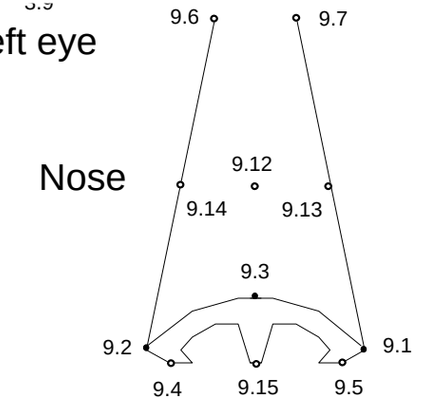
Teeth

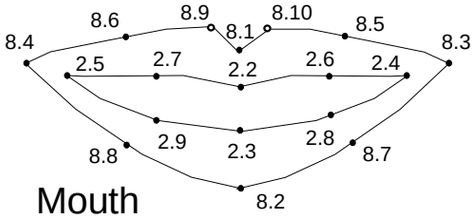


Tongue

Right eye

Left eye





Mouth

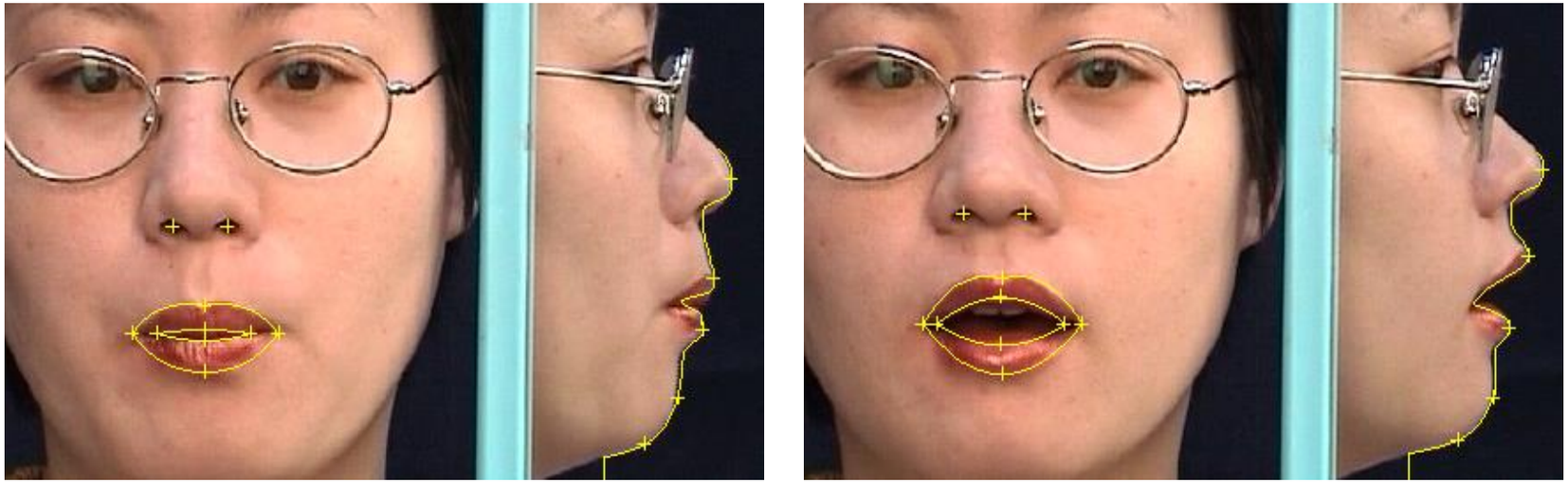
- Feature points affected by FAPs
- Other feature points

MPEG-4 Facial Animation Parameters (FAP)

FAP#	名 称	FAP#	名 称	FAP#	名 称
3	open_jaw	13	raise_r_cornerlip	53	stretch_l_cornerlip_o
4	lower_t_lip	14	thrust_jaw	54	stretch_r_conerlip_o
5	raise_b_midlip	16	push_b_lip	55	lower_t_lip_lm_o
6	stretch_l_cornerlip	17	push_t_lip	56	lower_t_lip_rm_o
7	stretch_r_conerlip	44	raise_tongue_tip	57	raise_b_lip_lm_o
8	lower_t_lip_lm	45	thrust_tongue_tip	58	raise_b_lip_rm_o
9	lower_t_lip_rm	46	raise_tongue	59	raise_l_cornerlip_o
10	raise_b_lip_lm	47	tongue_roll	60	raise_r_cornerlip_o
11	raise_b_lip_rm	51	lower_t_lip_o		
12	raise_l_cornerlip	52	raise_b_midlip_o		

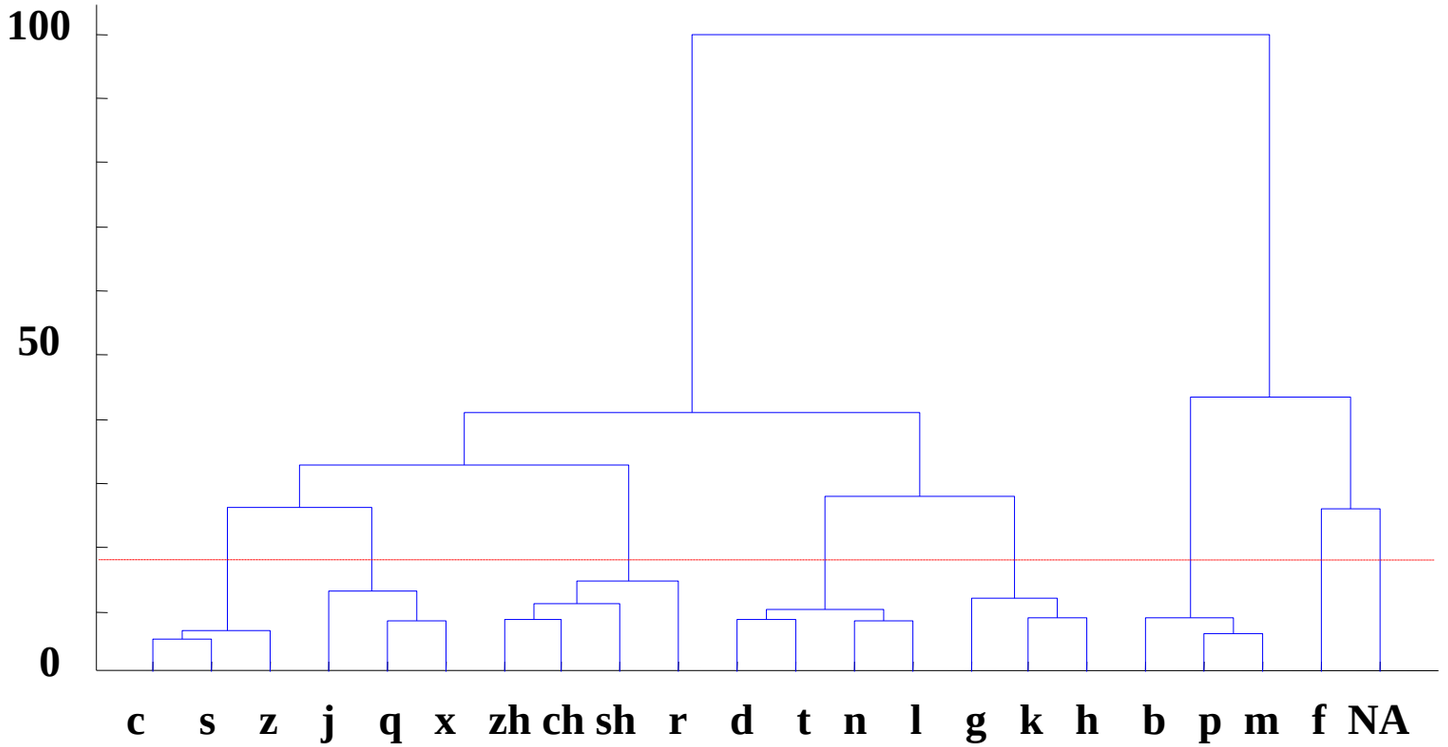
FAP Demo: Expression Editor

静态视位参数估计及汉语视位分类



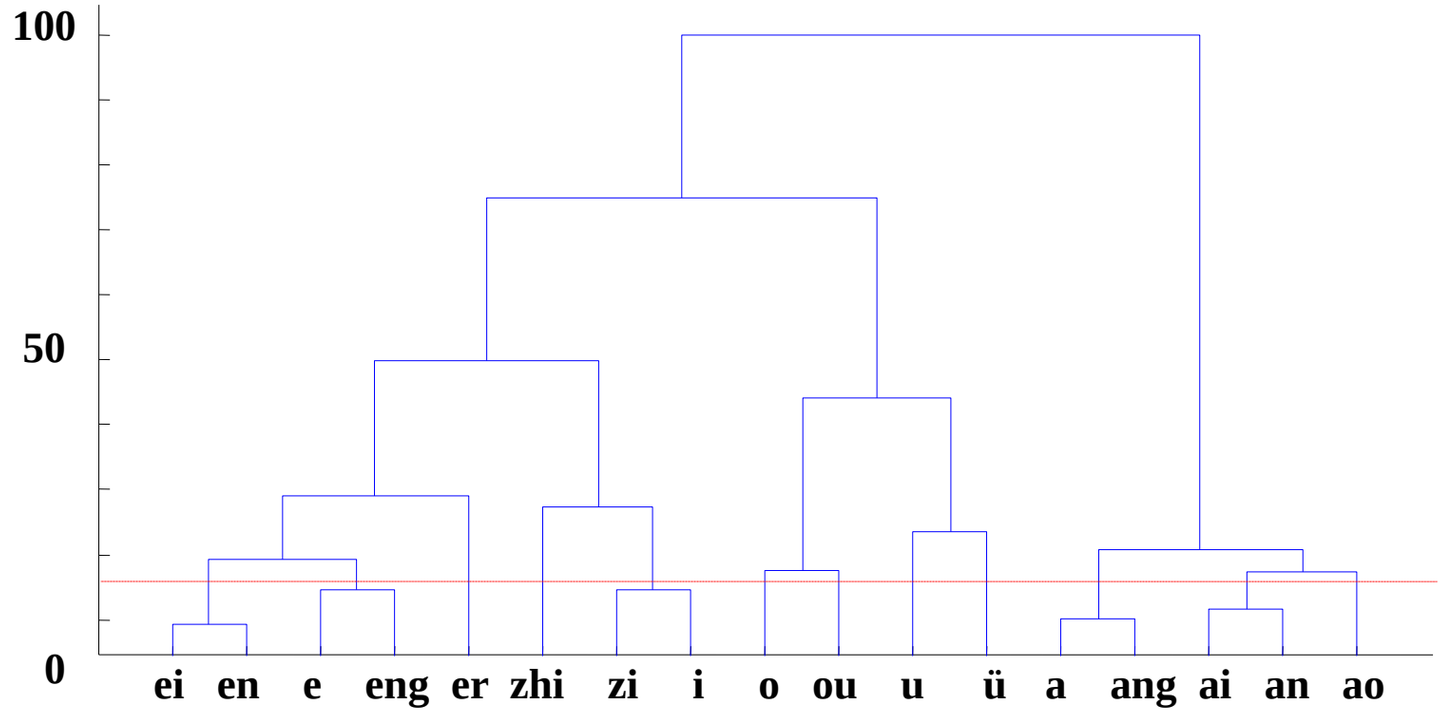
静态视位参数估计及汉语视位分类

- 声母视觉混淆树



静态视位参数估计及汉语视位分类

- 韵母视觉混淆树

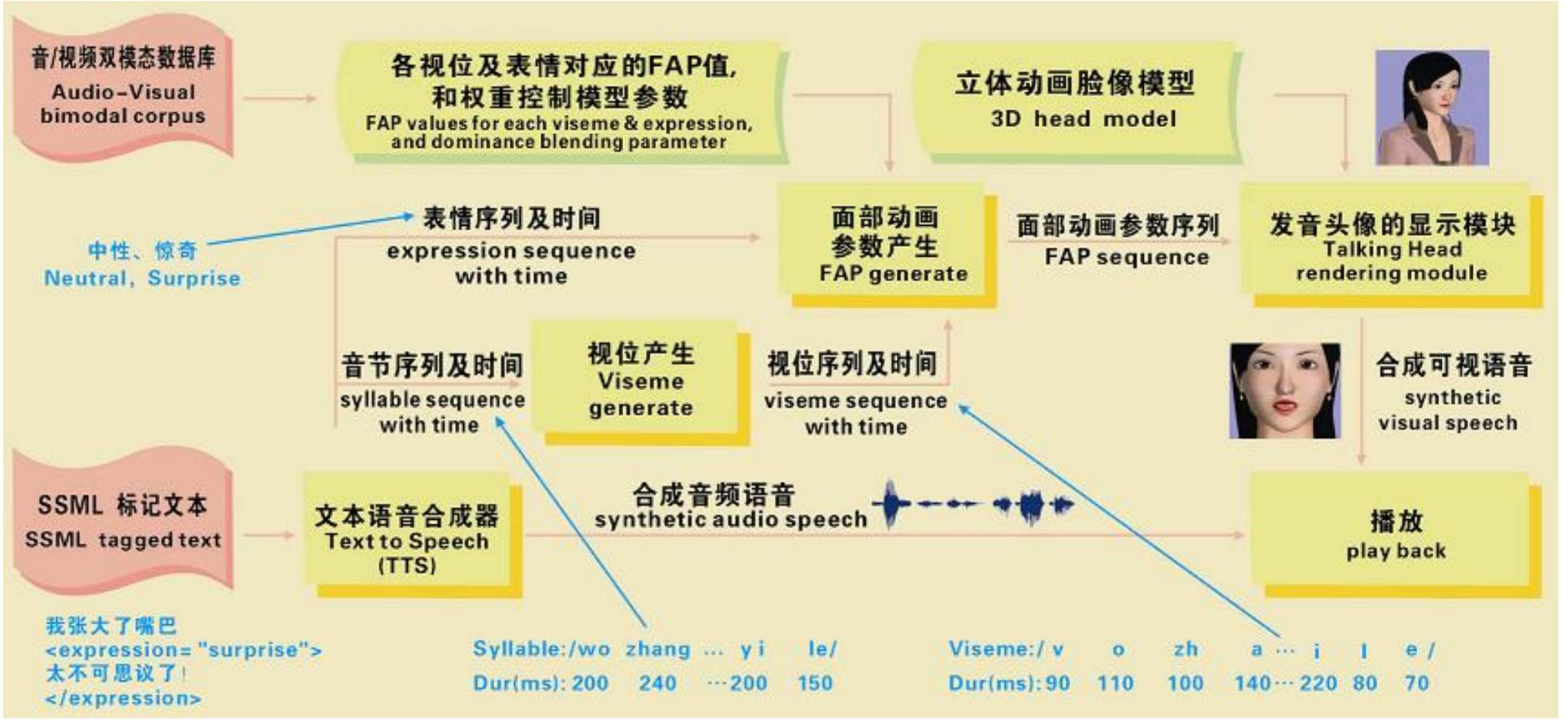


静态视位参数估计及汉语视位分类

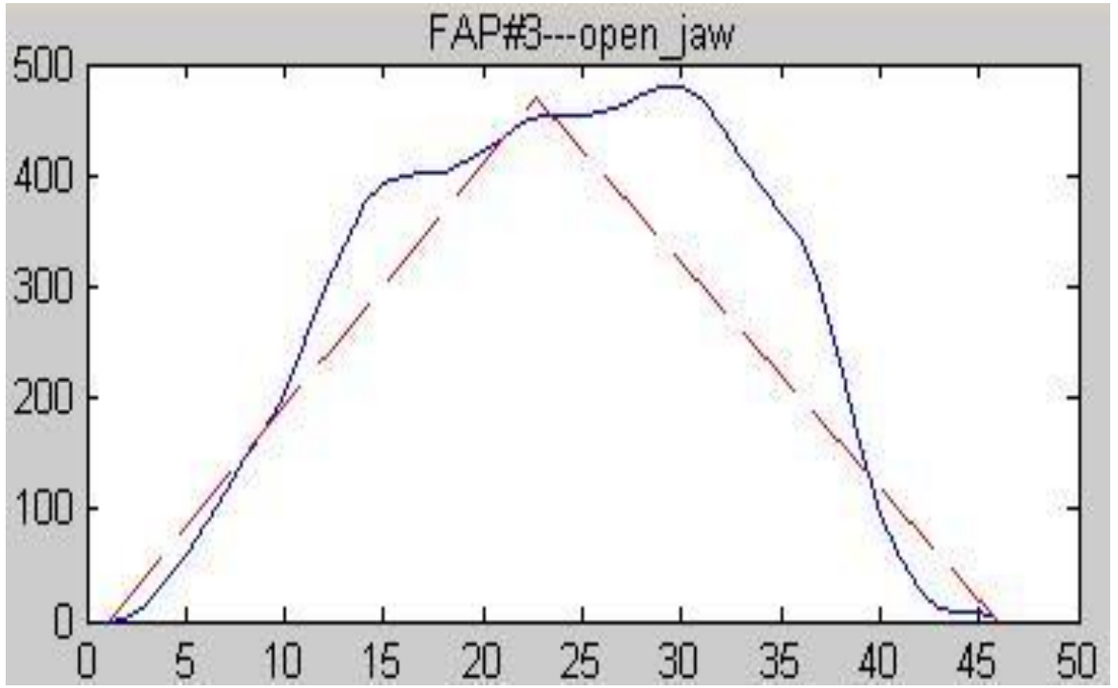
- 汉语基本静态视位分类

视位号	对应声韵母	视位号	对应声韵母	视位号	对应声韵母
0	NA(自然状态)	7	z, c, s	14	i (资韵)
1	b, p, m	8	a, ang	15	o
2	f	9	ai, an	16	ou
3	d, t, n, l	10	ao	17	u
4	g, k, h	11	e, eng	18	ü
5	j, q, x	12	ei, en	19	-i (知韵)
6	zh, ch, sh, r	13	er		

基于参数控制的可视语音合成



静态视位 V.S. 动态视位



发/o/音时下腭张开度变化曲线

基于权值融合的动态视位模型

基于加权混合的动态视位模型（Weight Blending Dynamic Viseme Model: WB-DVM）：

基本权值函数：

$$W = \alpha e^{\theta|\tau|^c}$$

其中 α 给出视位峰值处的权值幅度；θ 为权值衰减或增加的速度；$|\tau|$ 表示当前时刻到权值函数中心点时刻的时间距离；c 为一常数

基于权值融合的动态视位模型

第*i* 个视位*p* 参数的基本权值函数：

$$W_{ip} = \alpha_{ip} e^{-\theta_{ip(-)}|\tau|^{c_p}} \qquad \tau \geq 0$$

$$W_{ip} = \alpha_{ip} e^{-\theta_{ip(+)}|\tau|^{c_p}} \qquad \tau < 0$$

声母： $\tau = t_{si} - t_{iip} - t$

韵母： $\tau = t_{ci} - t_{ifp} - t$

t_{si}、t_{ci} 分别为语音段的起始时刻和中心时刻，t_{iip} 为从语音段的开始时刻到声母权值函数参数中心点的距离，t_{ifp} 为从语音段的中心时刻到韵母权值函数参数中心点的距离

基于权值融合的动态视位模型

左无声模型：

$$W_{lp} = \alpha_{lp} e^{\theta_{lp}|\tau|^{c_p}} \quad \tau \geq 0$$

$$W_{lp} = \alpha_{lp} e^{-\theta_{lp}|\tau|^{c_p}} \quad \tau < 0$$

$$\tau = t_{si} - t_{lp} - t$$

右无声模型：

$$W_{rp} = \alpha_{rp} e^{-\theta_{rp}|\tau|^{c_p}} \quad \tau \geq 0$$

$$W_{rp} = \alpha_{rp} e^{\theta_{rp}|\tau|^{c_p}} \quad \tau < 0$$

$$\tau = t_{ei} - t_{rp} - t$$

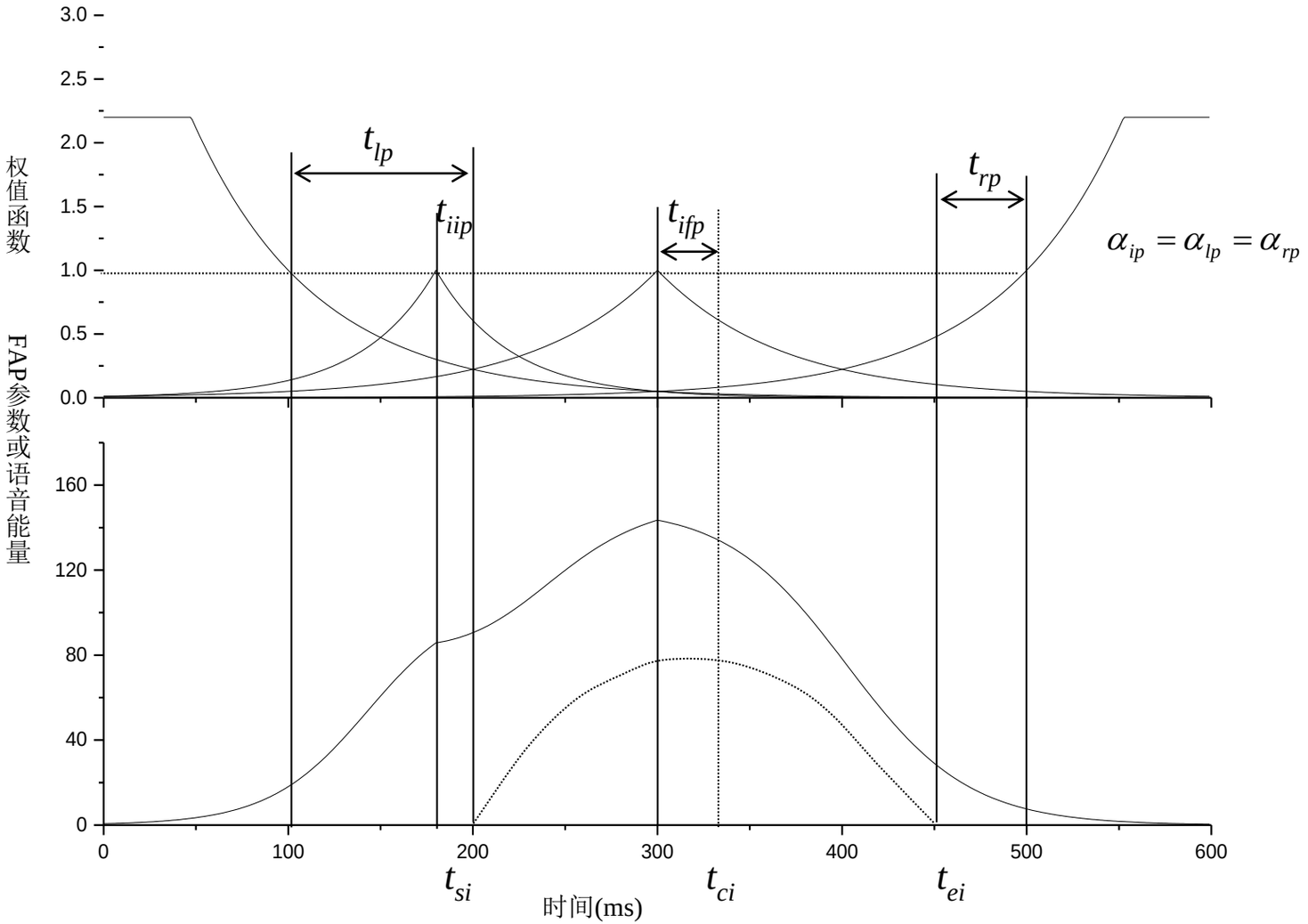
t_{si} 表示语音开始时刻；

t_{ei} 表示语音结束时刻；

t_{lp} 表示从语音起始时刻到左无声模型中心的时间距离；

t_{rp} 表示从语音结束时刻到右无声模型中心的时间距离。

基于权值融合的动态视位模型



动态视位模型中各个时刻之间关系的示意图

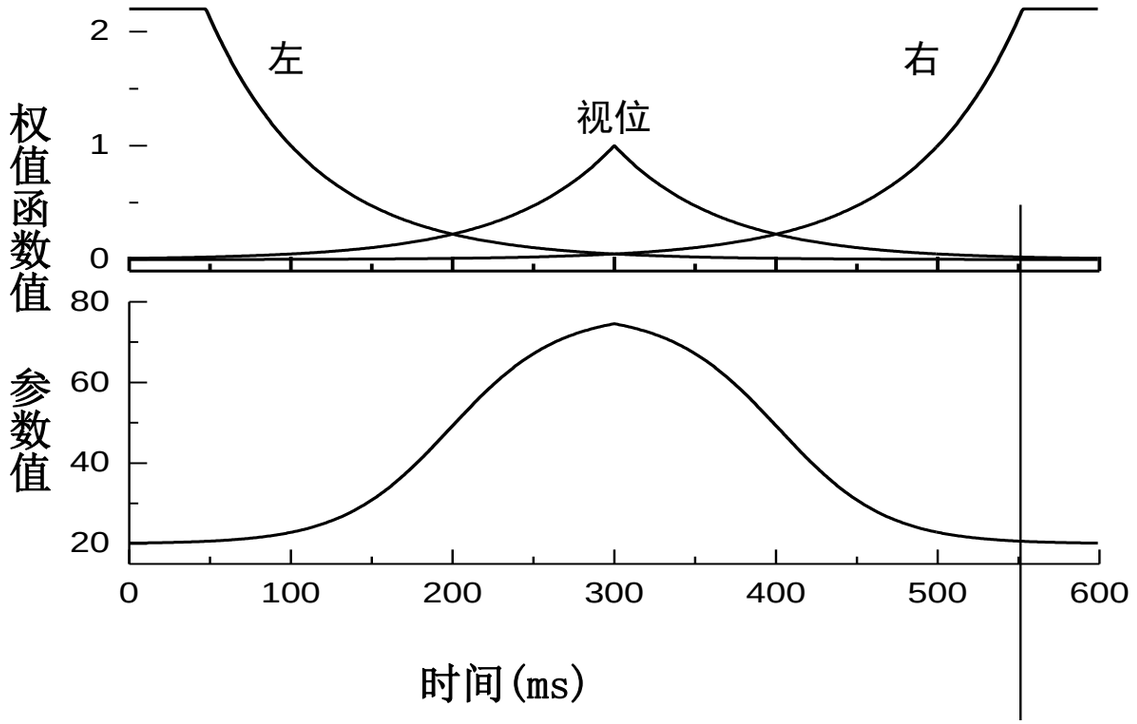
基于权值融合的动态视位模型

- 总的动态视位模型：

$$F_{ip}(t)=\frac{W_i(t)*T_{ip}+(W_l(t)+W_r(t))*T_{0p}}{W_i(t)+W_l(t)+W_r(t)}$$

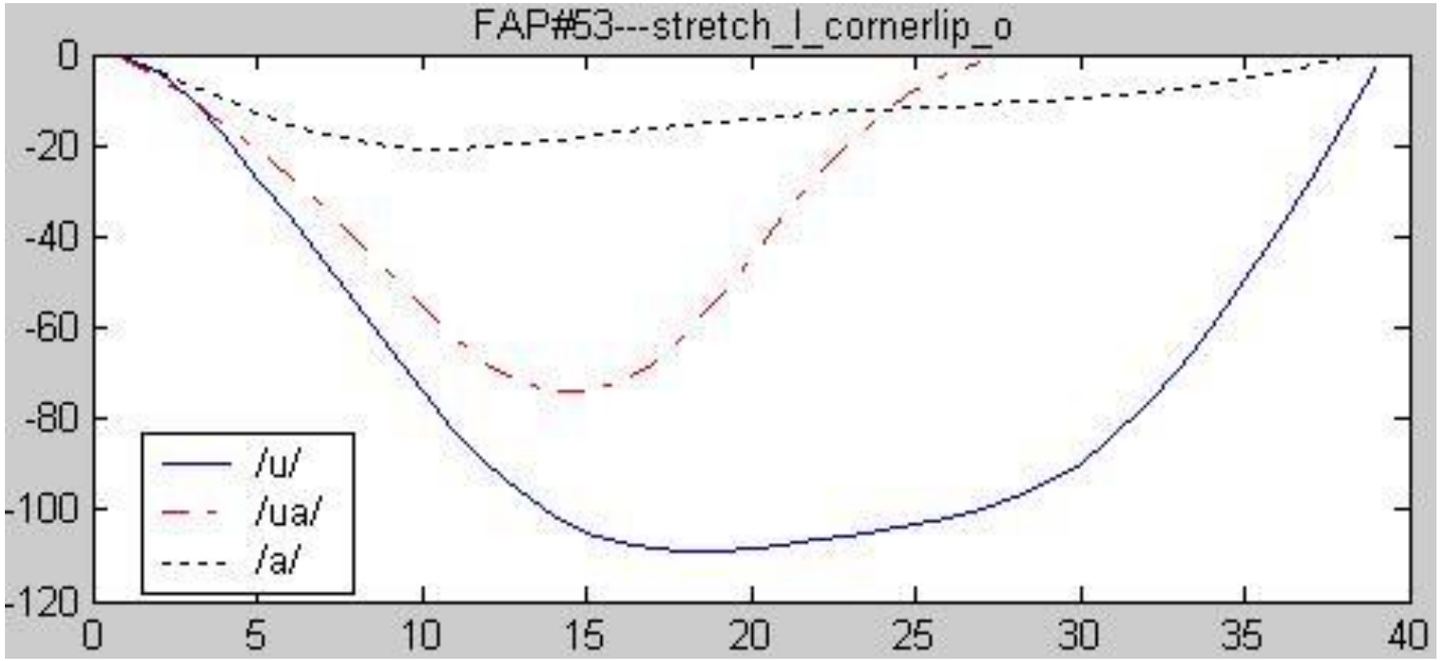
任意时刻视位参数生成

基于权值融合的动态视位模型



动态视位模型的参数变化

动态视位 与 协同发音 (co-articulation)



协同发音对视位参数的影响

基于权值融合的动态视位模型

连续语流中视位参数的生成：

$$F_p(t) = (\sum_{i=1}^{I}(W_i(t) \times T_{ip})) / (\sum_{i=1}^{I} W_i(t))$$

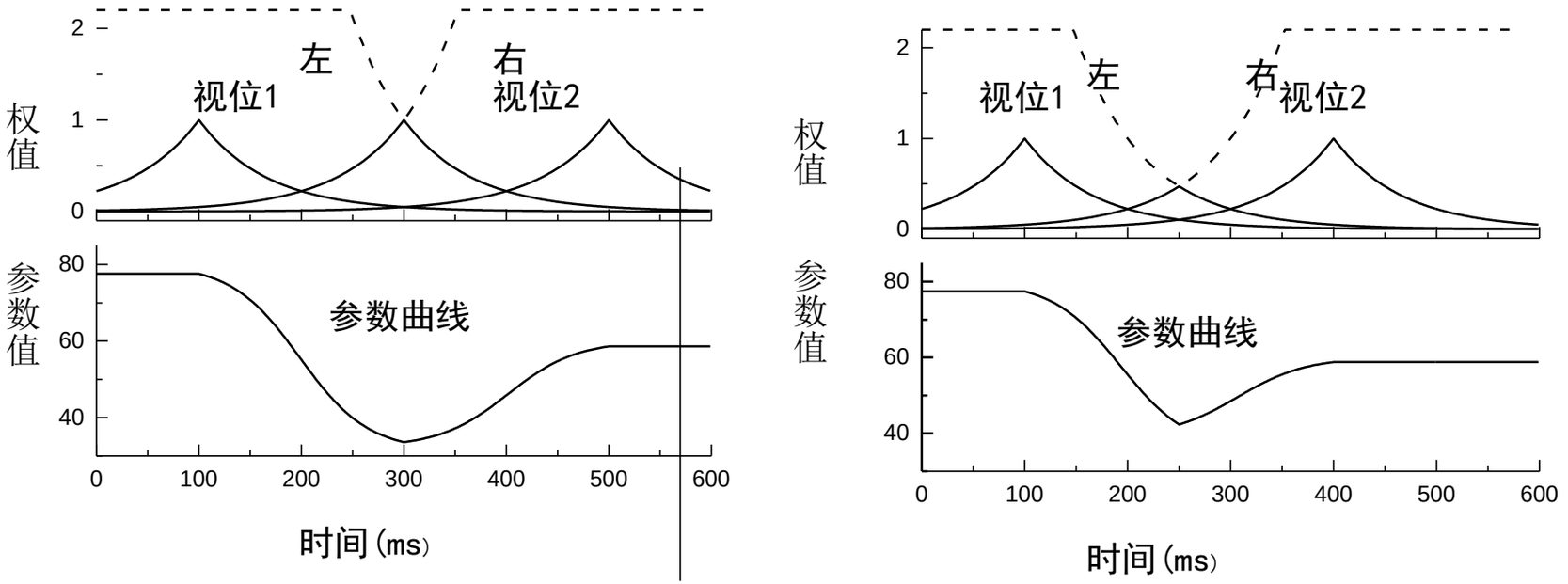
I 是协同发音所考虑的音位个数

特殊视位的特殊混合规则：

$$W_{ip} = W_{ip0} + (C\sum_{k \neq i} W_{kp} - W_{ip0})(1 - \frac{|\tau|}{T})$$

基于权值融合的动态视位模型

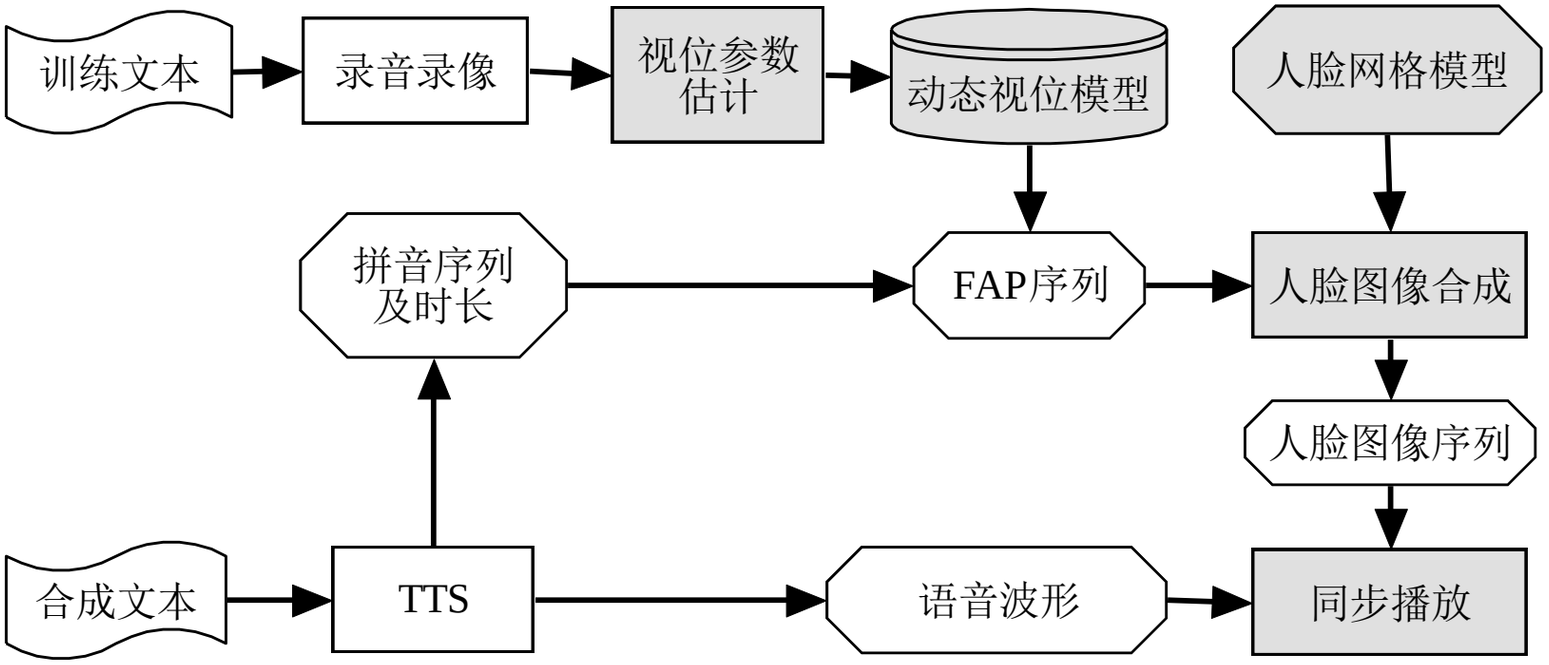
连续语流中的无声模型由前一个视位的右无声模型和后一个视位的左无声模型交叉形成（实线部分），不同停顿有不同的参数过渡曲线。



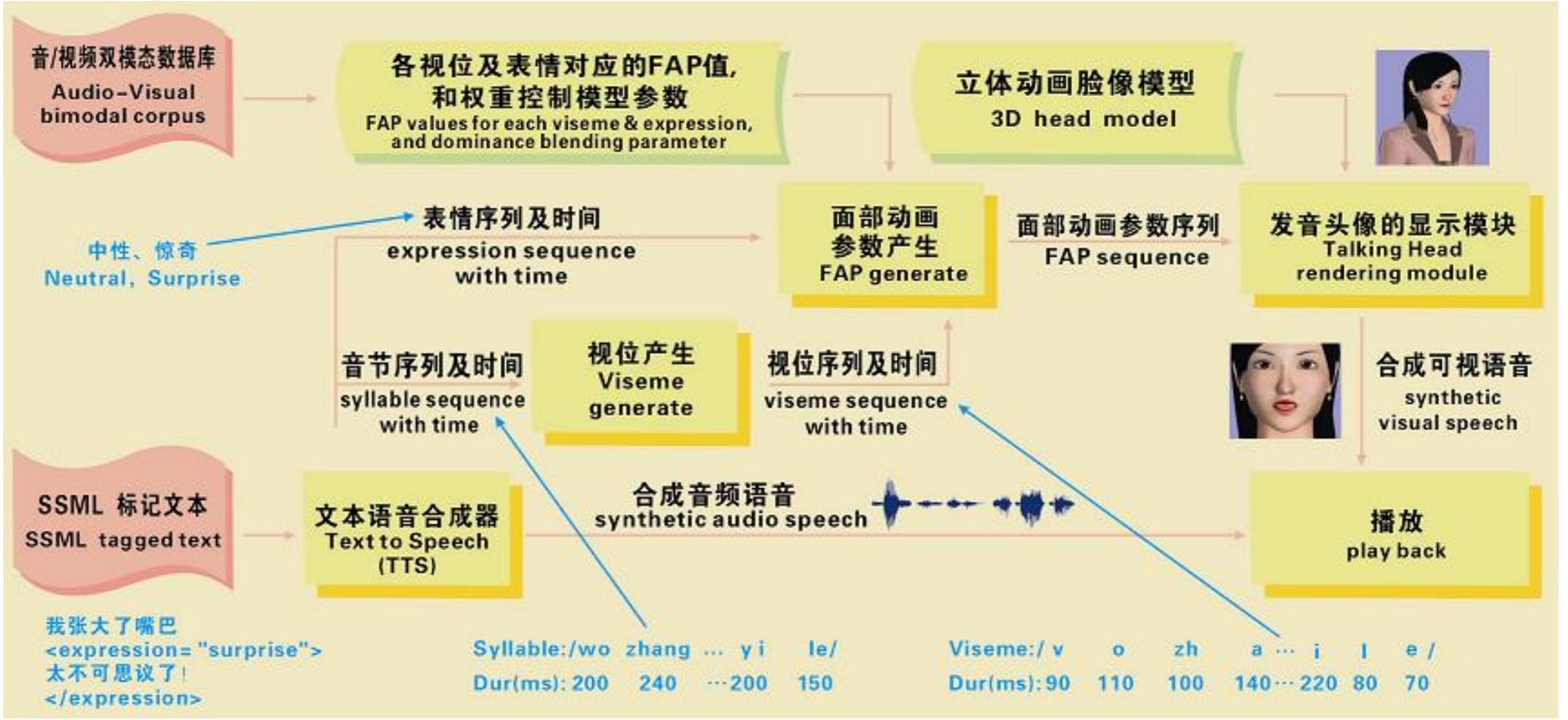
语流中不同停顿时间的参数变化

基于参数控制的可视语音合成

系统框图



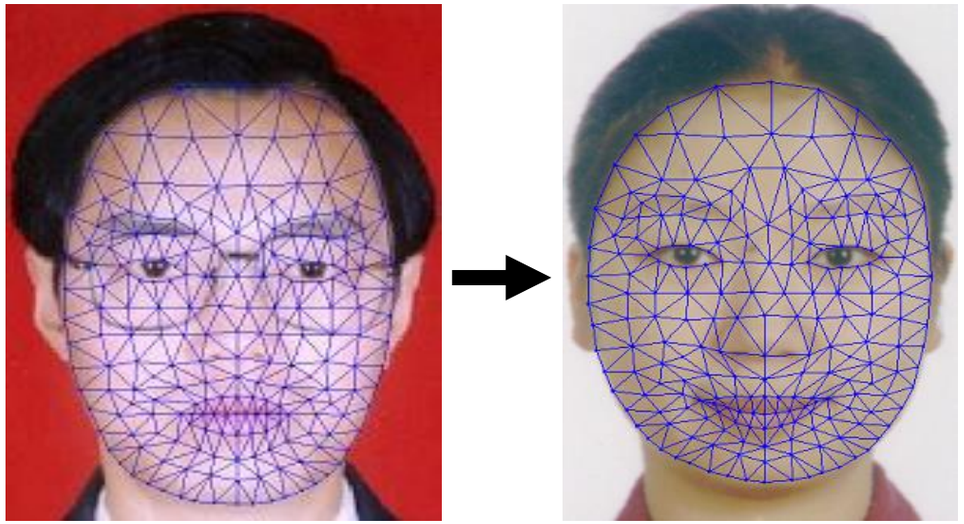
基于参数控制的可视语音合成



参数控制的可视语音合成

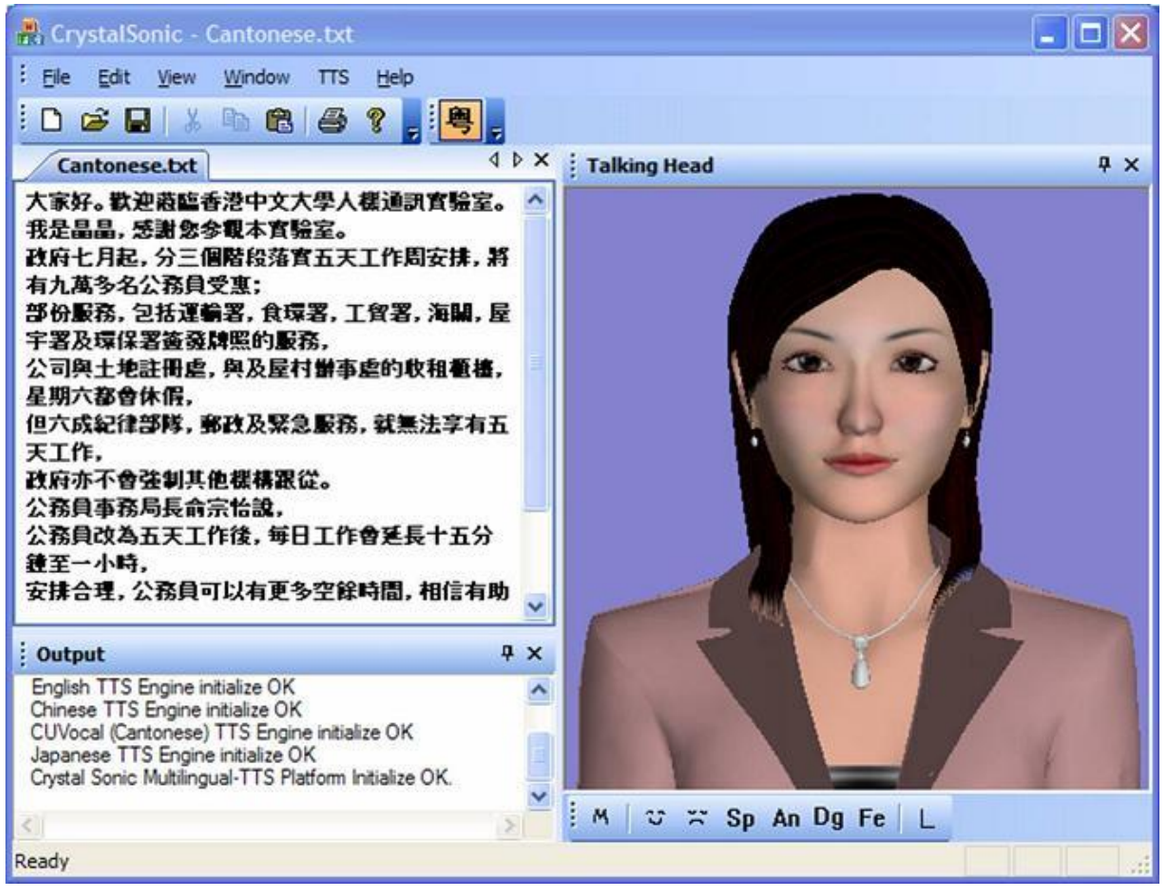
人脸模型及实现技巧：

- ✓2D人脸模型(219个顶点；362个三角形；581条边);
- ✓图像变形：双线性插值；
- ✓小幅度的头部运动；
- ✓眼珠的左右转动；
- ✓不定时的眨眼；
- ✓ 快速更换人脸模型.



几分钟

参数控制的可视语音合成



主要内容

- 视位（Viseme）的概念
- 基于参数控制的可视语音合成
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- **基于数据驱动的可视语音合成**
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- 基于深度学习技术的可视语音合成

基于数据驱动的可视语音合成

- 视位的视频序列表征
  - 直接用小段视频序列表示视位
  - 包括了可视发音器官的形状、位置，以及动作

Chen, T. (2001). *Audiovisual speech processing.* IEEE Signal Processing Magazine, 9–31.

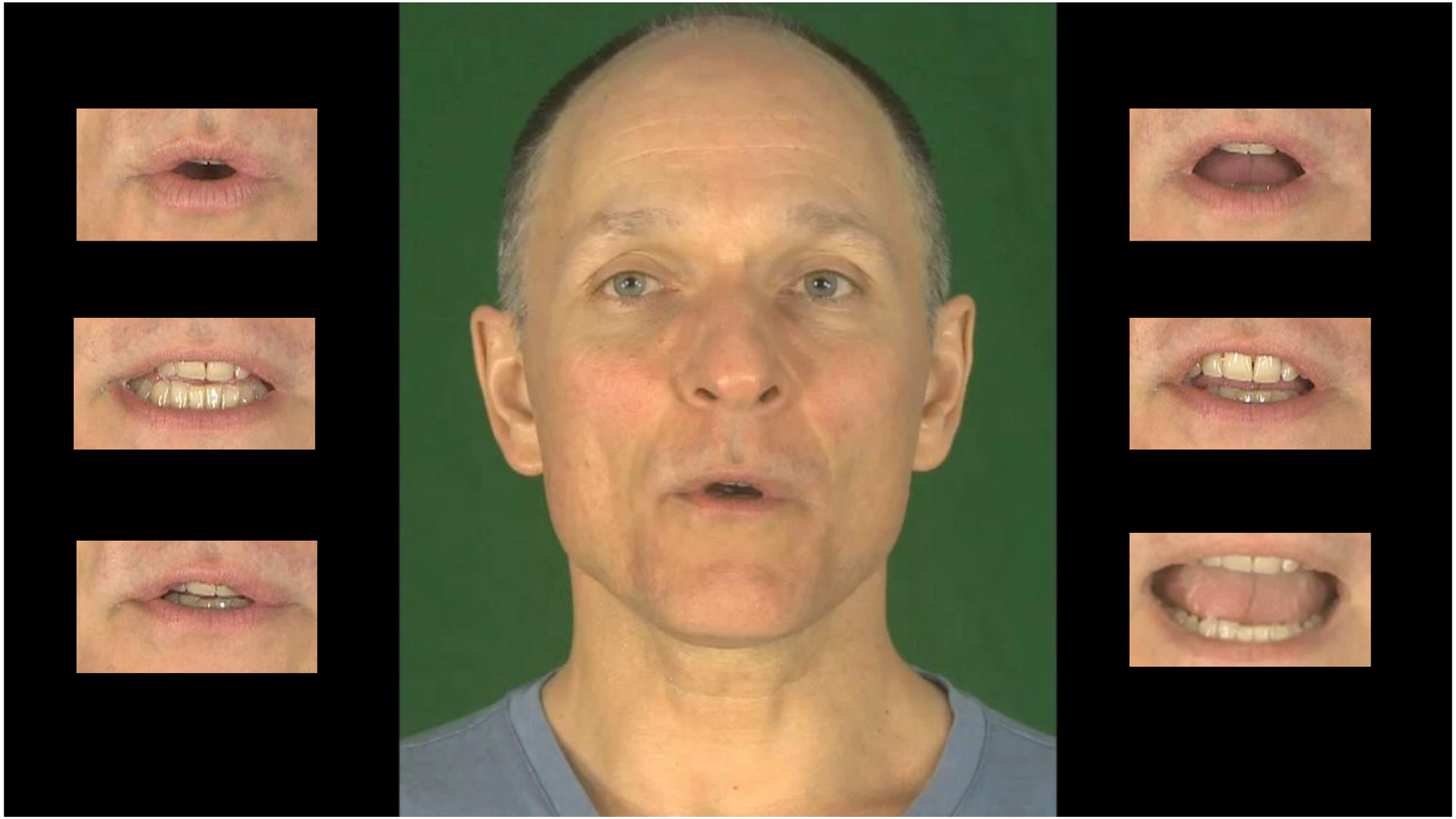
基于数据驱动的可视语音合成

- 语音中的协同发音(Co-articulation)现象

(a) wu-de-xiu

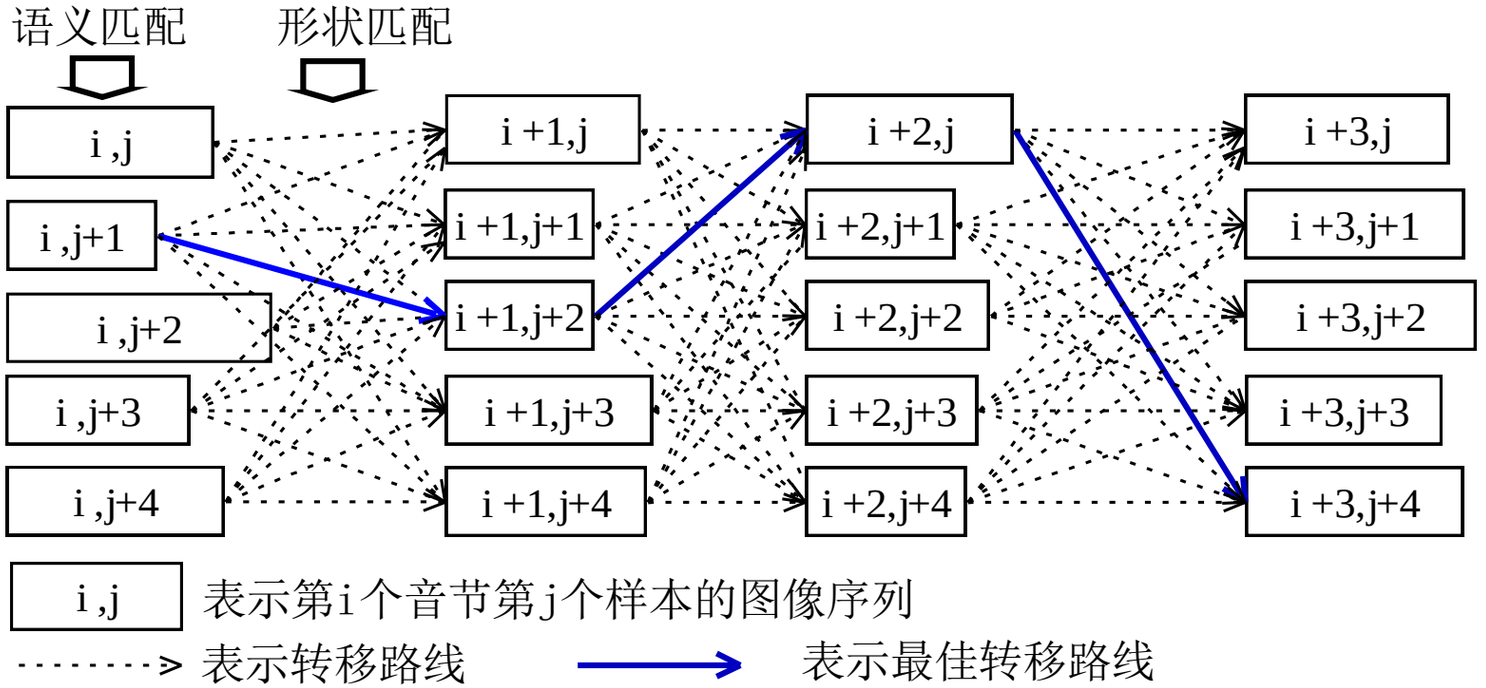
(b) da-de-xiong

Co-Articulation is Hard



/k/

基于数据驱动的可视语音合成



语义与形状相结合的基元选取

基元选取代价函数：

$$Cost = \alpha \sum_{i=1}^{N} VD_{i,i'} + (1-\alpha) \sum_{i=1}^{N-1} SD_{i',i'+1}$$

形状距离：

$$SD_{i',i'+1} = \frac{\left|H_{i'} - H_{i'+1}\right|}{Max(H) - Min(H)} + \frac{\left|W_{i'} - W_{i'+1}\right|}{Max(W) - Min(W)}$$

H 嘴唇的高度，W 嘴唇的宽度

基于硬度因子和视觉距离的协同发音模型：

$$VD_{i,j} = H_{-2} \cdot (1 - H_{-1}) \cdot D_{i-2,j-2} + H_{-1} \cdot D_{i-1,j-1} + H_{+1} \cdot D_{i+1,j+1} + H_{+2} \cdot (1 - H_{+1}) \cdot D_{i+2,j+2}$$

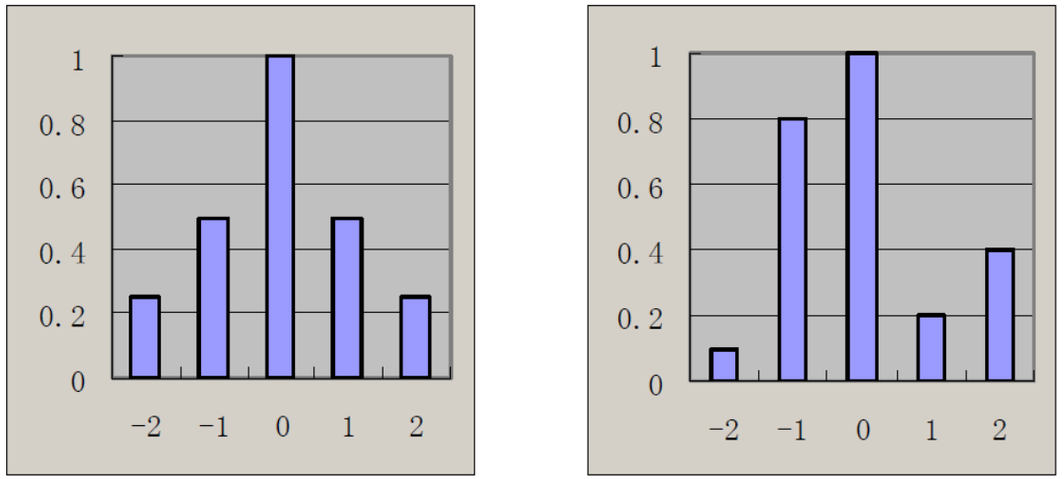
其中 $VD_{i,j}$ 表示韵母 j 与韵母 i 的视觉距离；

$H_t = Max(H_{t,i}, H_{t,j})$ t={-2,-1,+1,+2} 给出了韵母 j 和韵母 i 的前一个韵母、前一个声母、后一个声母及后一个韵母的硬度因子($0 \le H_{t,i}, H_{t,j} \le 1$)。

$D_{i-2,j-2}$, $D_{i-1,j-1}$, $D_{i+1,j+1}$, $D_{i+2,j+2}$ 给出了两个不同发音环境中前一个韵母、前一个声母、后一个声母及后一个韵母的视觉距离，这些距离可从前面所述的汉语声韵母视觉混淆树中得到。

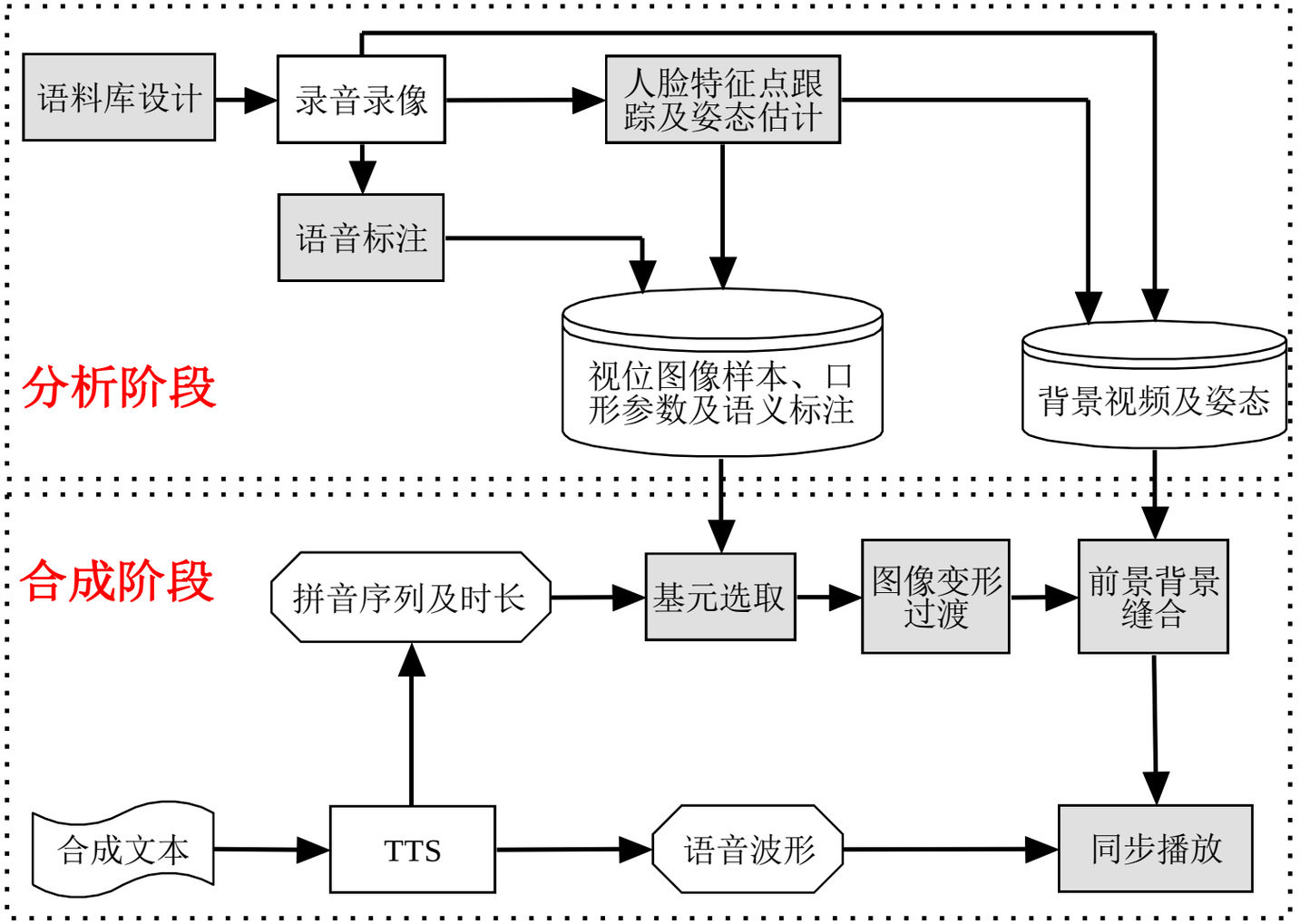
基于硬度因子和视觉距离的协同发音模型：

$$VD_{i,j} = H_{-2} \cdot (1 - H_{-1}) \cdot D_{i-2,j-2} + H_{-1} \cdot D_{i-1,j-1} + H_{+1} \cdot D_{i+1,j+1} + H_{+2} \cdot (1 - H_{+1}) \cdot D_{i+2,j+2}$$



不同硬度因子所造成的视觉距离权值的变化
左图：$H_{-2}, H_{-1}, H_{+1}, H_{+2}$均为0.5
右图：H_{-2}, H_{+2}为0.5, H_{-1}为0.8, H_{+1}为0.2

基于数据驱动的可视语音合成

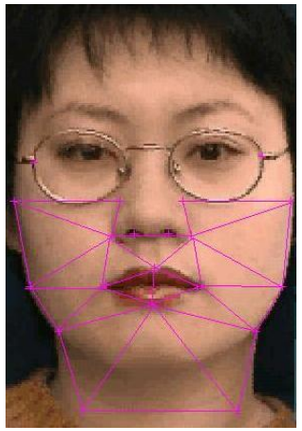


基于数据驱动的可视语音合成系统

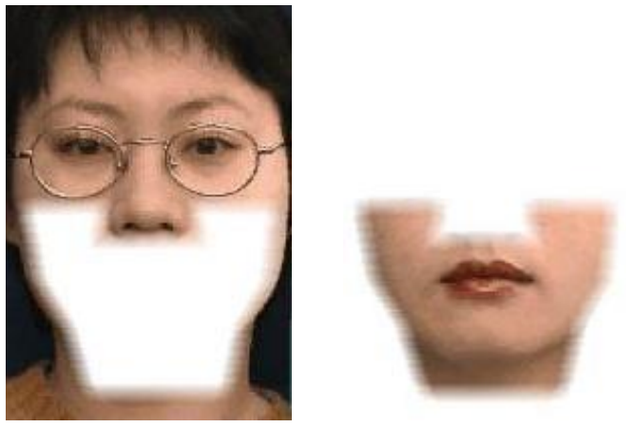
图像的平滑过渡变形与叠加合成

■ A B

图像变形平滑过渡(上: 原始图；下：平滑结果； α =0.3)



图像变形的
人脸三角化结构



人脸背景和前景的叠加合成

基于数据驱动的可视语音合成

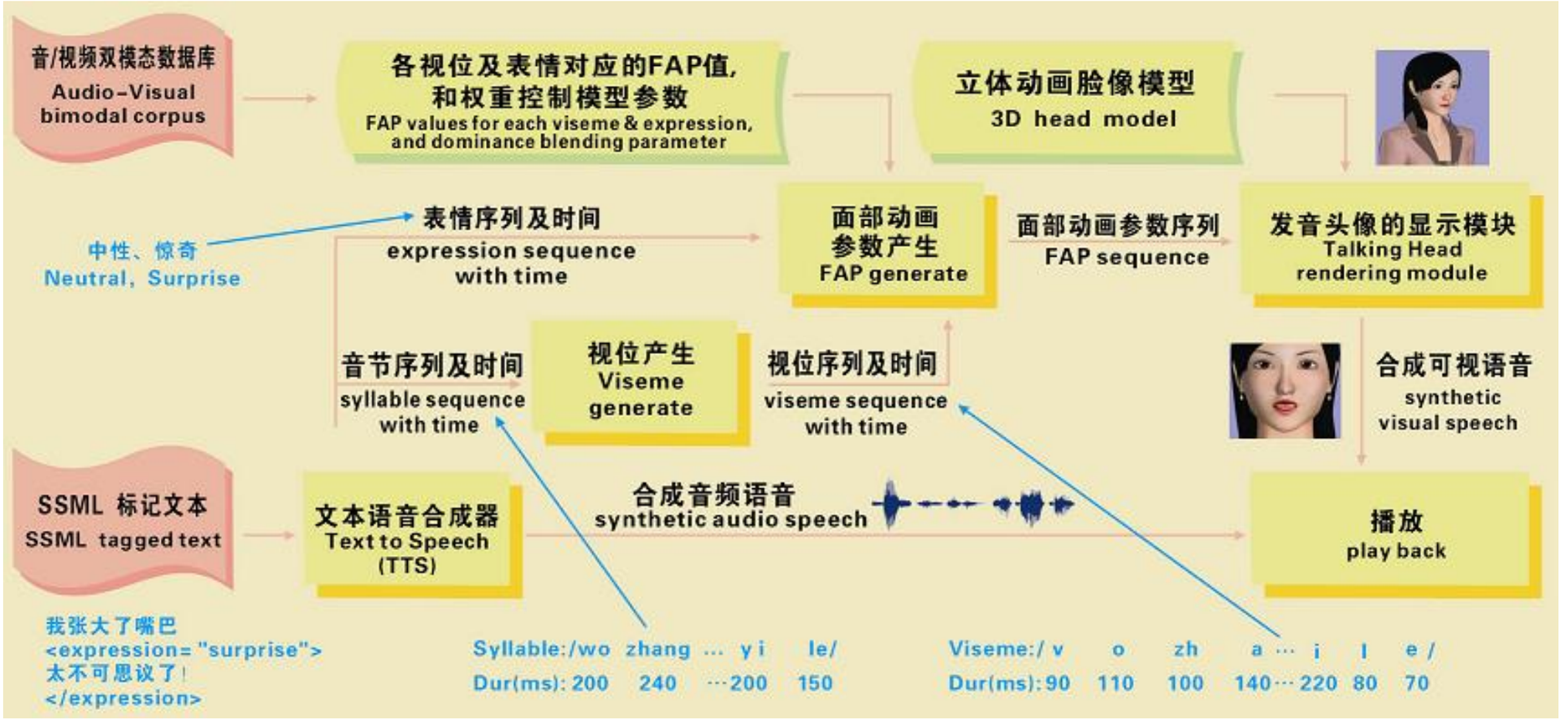
- **Microsoft Bing (必应) Dict**

http://cn.bing.com/dict/search?q=proximity&FORM=BDVSP6

主要内容

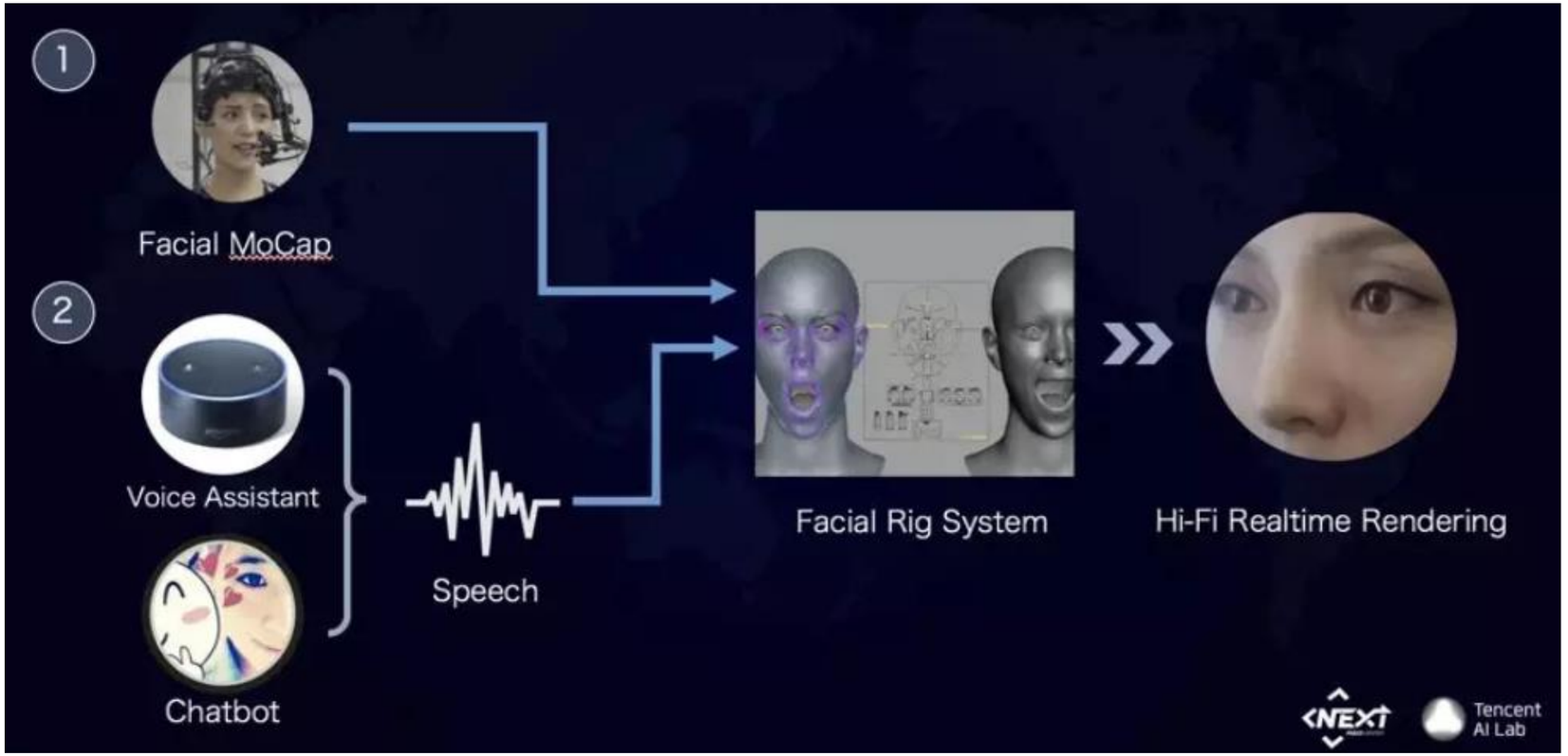
- 视位（Viseme）的概念
- 基于参数控制的可视语音合成
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- 基于数据驱动的可视语音合成
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- 基于深度学习技术的可视语音合成

Text-To-Visual-Speech, TTVS

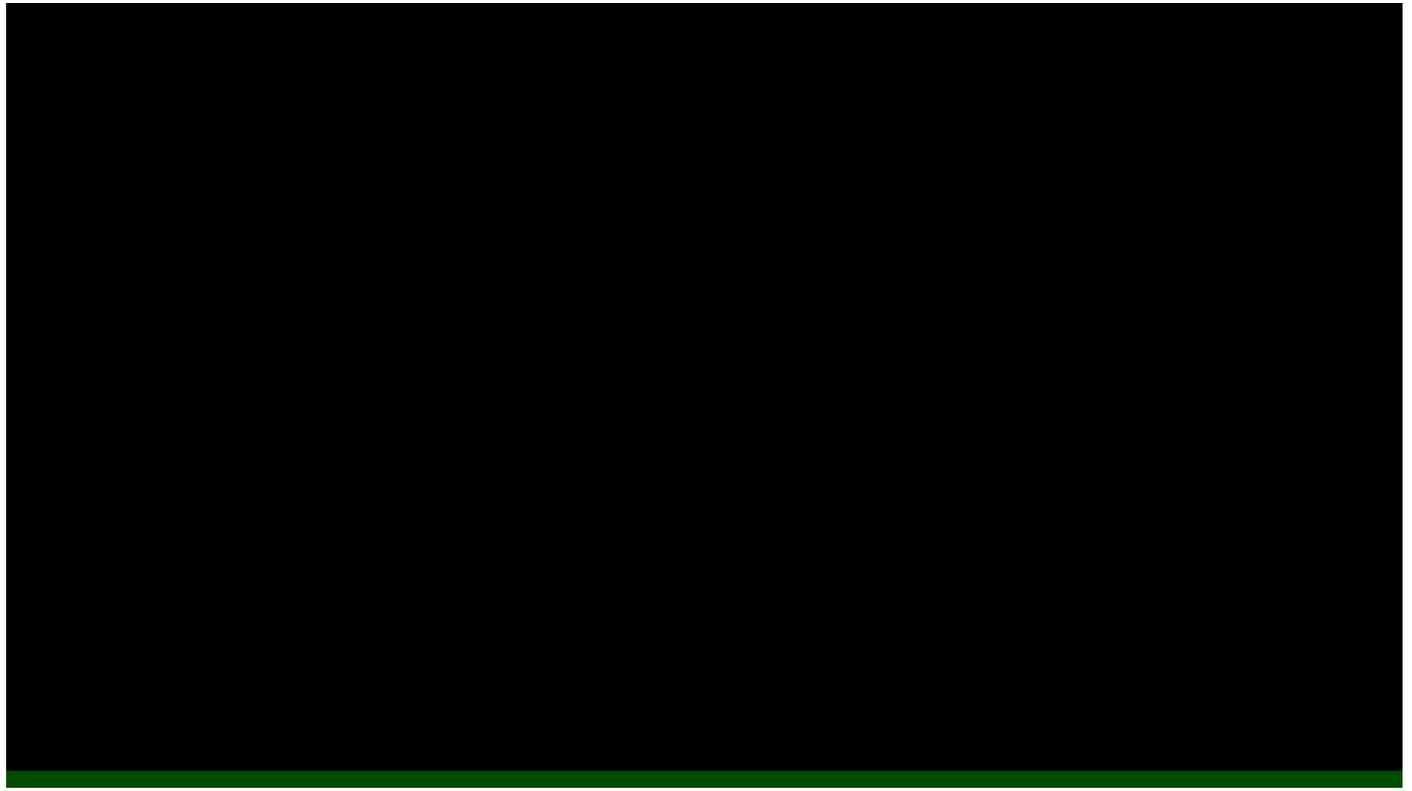


Speech Driven Talking Avatar

- **Audio Driven Facial Animation (ADFA)**



ObamaNet: Photo-realistic lip-sync from text



Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, Yoshua Bengio,
"ObamaNet: Photo-realistic lip-sync from text", NIPS 2017. [https://arxiv.org/abs/1801.01442]

ObamaNet: Photo-realistic lip-sync from text

Abstract

We present **ObamaNet**, the first architecture that takes any text as input and generates both the corresponding speech and synchronized photo-realistic lip-sync videos. Contrary to other published lip-sync approaches, ours is only composed of fully trainable neural modules and does not rely on any traditional computer graphics methods. More precisely, we use three main modules: a text-to-speech network based on **Char2Wav**, a time-delayed LSTM to generate mouth-keypoints synced to the audio, and a network based on **Pix2Pix** to generate the video frames conditioned on the keypoints.

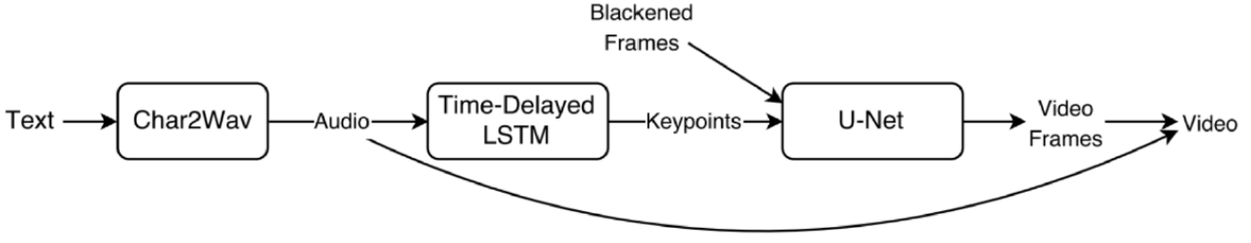


Figure 1: Flow diagram of our generation system.

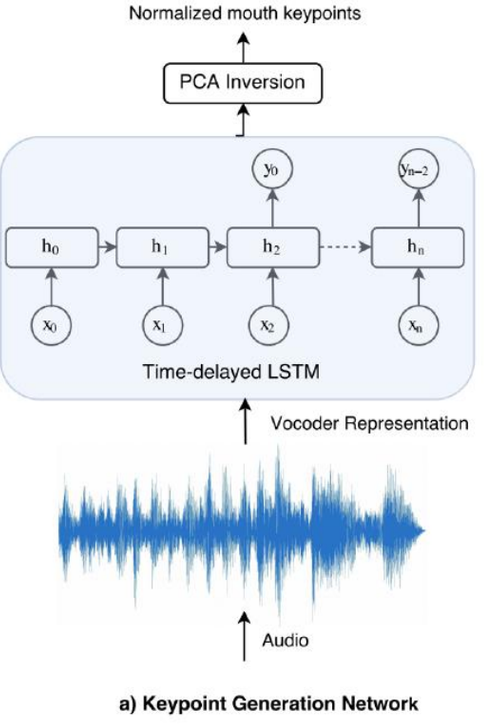


Figure 2: Keypoint Generation Network

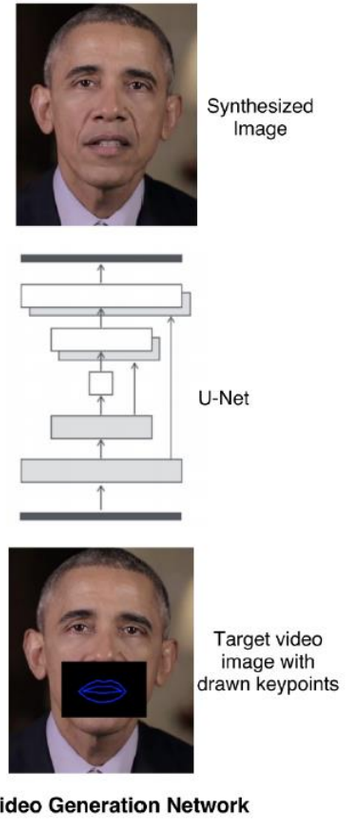


Figure 3: Video Generation Network

Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, Yoshua Bengio, "ObamaNet: Photo-realistic lip-sync from text", NIPS 2017. [https://arxiv.org/abs/1801.01442]

“视位”的参数化描述: Facial Keypoints

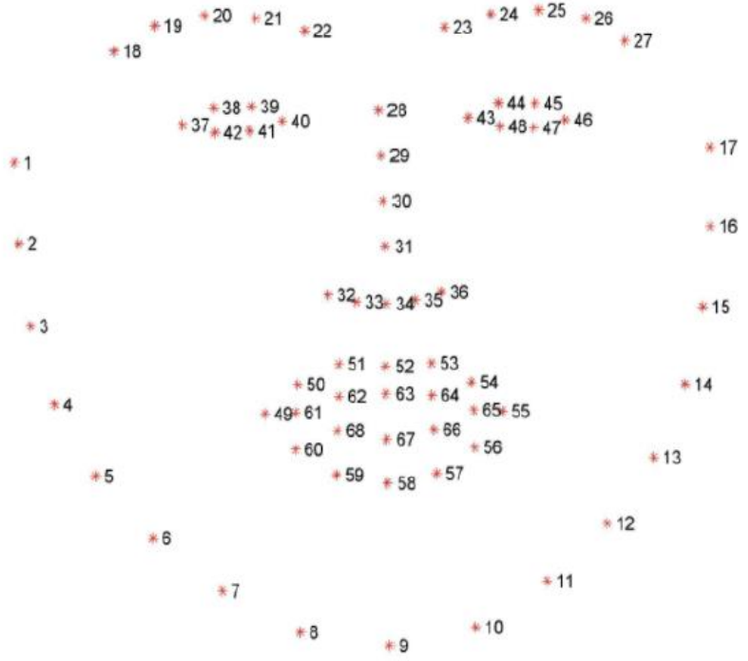


Figure 4: The 68 facial keypoints

Keypoint Generation The data required for the keypoint generation component is a representation of audio for input, and a representation of mouth shape for the output.

To compute the mouth shape representation, we extract 68 facial keypoints from each frame of the video. We used the publicly available dlib facial landmark detector to detect the 68 keypoints from the image. Sample annotations performed by the detector are shown in Figure 3.

These keypoints are highly dependent on the face location, face size, in-plane and out-of-plane face rotation. These variances are due to varying zoom-levels of the camera, distance between camera and speaker, and the natural head-motion of the speaker. In an effort to remove these variances, we first mean-normalize the 68 keypoints with the center of the mouth. This converts the 68 keypoints into vectors originating from the center of the mouth, thereby making it invariant to the face location.

To remove the in-plane rotation caused due to head motion, we project the keypoints into a horizontal axis using rotation of axes.

We make the keypoints invariant to face size, by dividing the keypoints by the norm of the 68 vectors from the center of the mouth, which serves as an approximation of face size.

Finally, we apply PCA to de-correlate the 20 normalized keypoints (40-D vector). We noticed that the first 5 PCA-coefficients capture >98% variability in the data.

语音到Facial Keypoints参数的生成

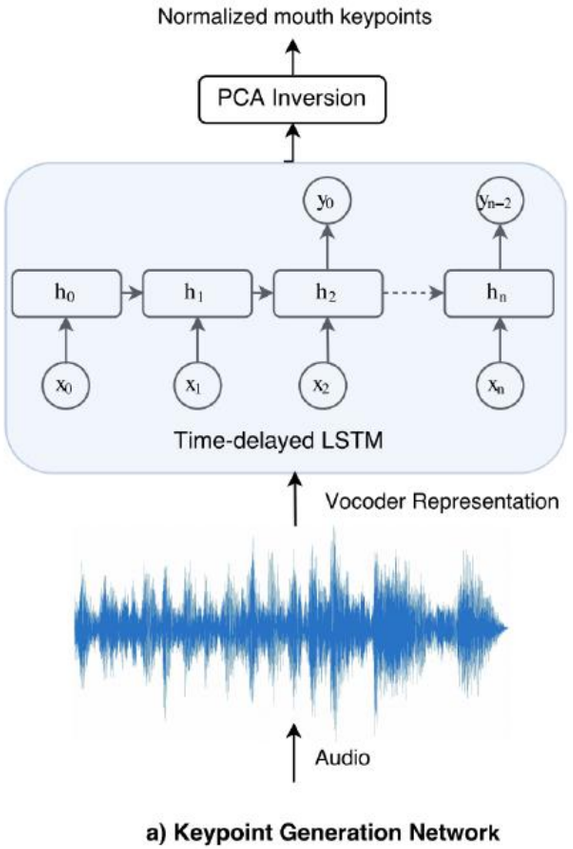


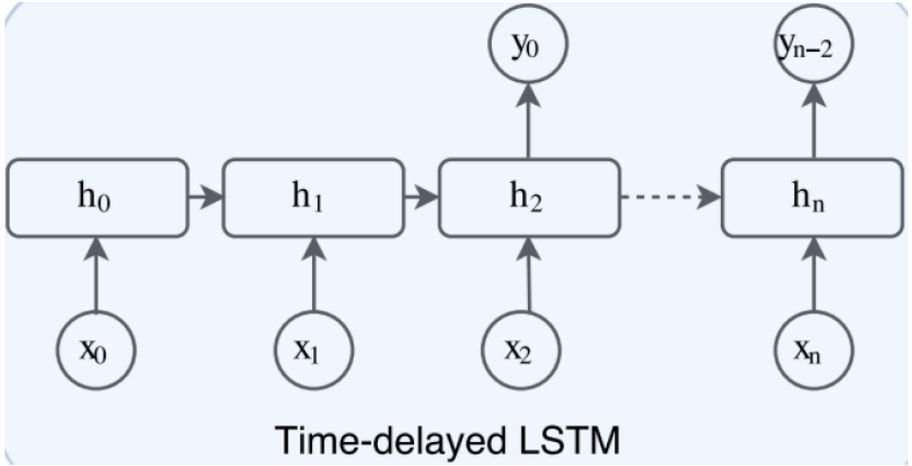
Figure 2: Keypoint Generation Network

- **Input: audio features**
  - Vocoder frames extracted from 16kHz audo
  - WORLD vocoder used
- **Output: mouth shape representation: Facial Keypoints**
  - For each frame of the face video, detect / extract 68 facial keypoints using dlib
  - Normalizations
    - Face location: mean normalization
    - Face rotation (in-plane): projection into a horizontal axis using rotation of axes
    - Face size: divide by the norm of 68 vectors
  - Apply PCA
    - Reduce dimensionality and de-correlate the 20 normalized keypoints (40-D vector)
    - The first 5 coefficients capture >98% variability
- **Network: Time-delayed LSTM**

Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, Yoshua Bengio, "ObamaNet: Photo-realistic lip-sync from text", NIPS 2017. [https://arxiv.org/abs/1801.01442]

语音到Facial Keypoints参数的生成

- **Network: Time-delayed LSTM**
  - Mouth moves before you say something
    - I.e., by the time Obama says *Uhhh*, his mouth is already open
    - Hence it is not enough to condition your mouth shape (*only*) on the past audio input
  - Simpler way to introduce a short future context is to add a time delay to the output
    - Shift the network output forward as *target delay*
    - E.g., target delay = 2



Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, Yoshua Bengio, "ObamaNet: Photo-realistic lip-sync from text", NIPS 2017. [https://arxiv.org/abs/1801.01442]

视频动画生成：Video Generation

- **Input**
  - Face image cropped around the mouth area and annotated with the outline
  - Crop the mouth area using a bounding box around the keypoints and draw the outline with OpenCV
- **Output**
  - Complete face image with *in-painted* mouth area
  - Denormalizations to ensure the rendered mouth to be compatible with the face in the target video
- **Network: Pix2pix (Isola etal. 2016)**
  - Train objective: L1-loss in pixel space (w/o GAN objective)

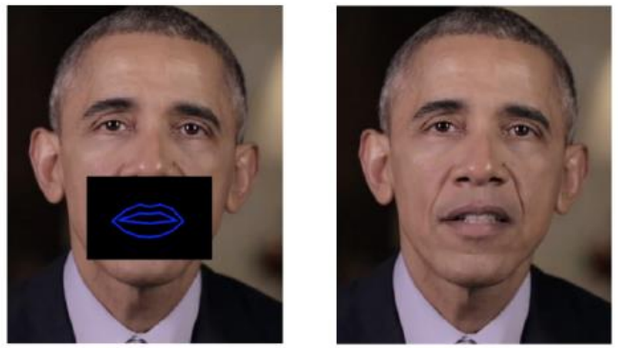


Figure 5: Sample input-output pair for the in-painting network

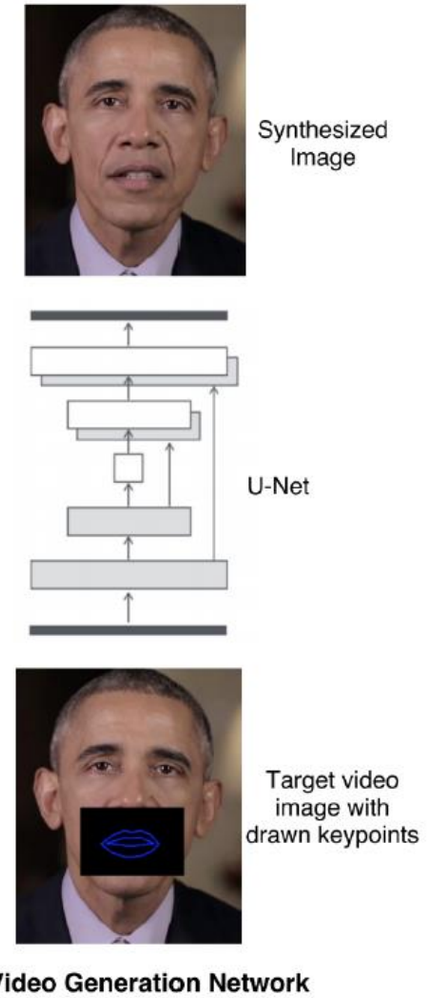


Figure 3: Video Generation Network

Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, "ObamaNet: Photo-realistic lip-sync from text", NIPS 2017. [https://arxi

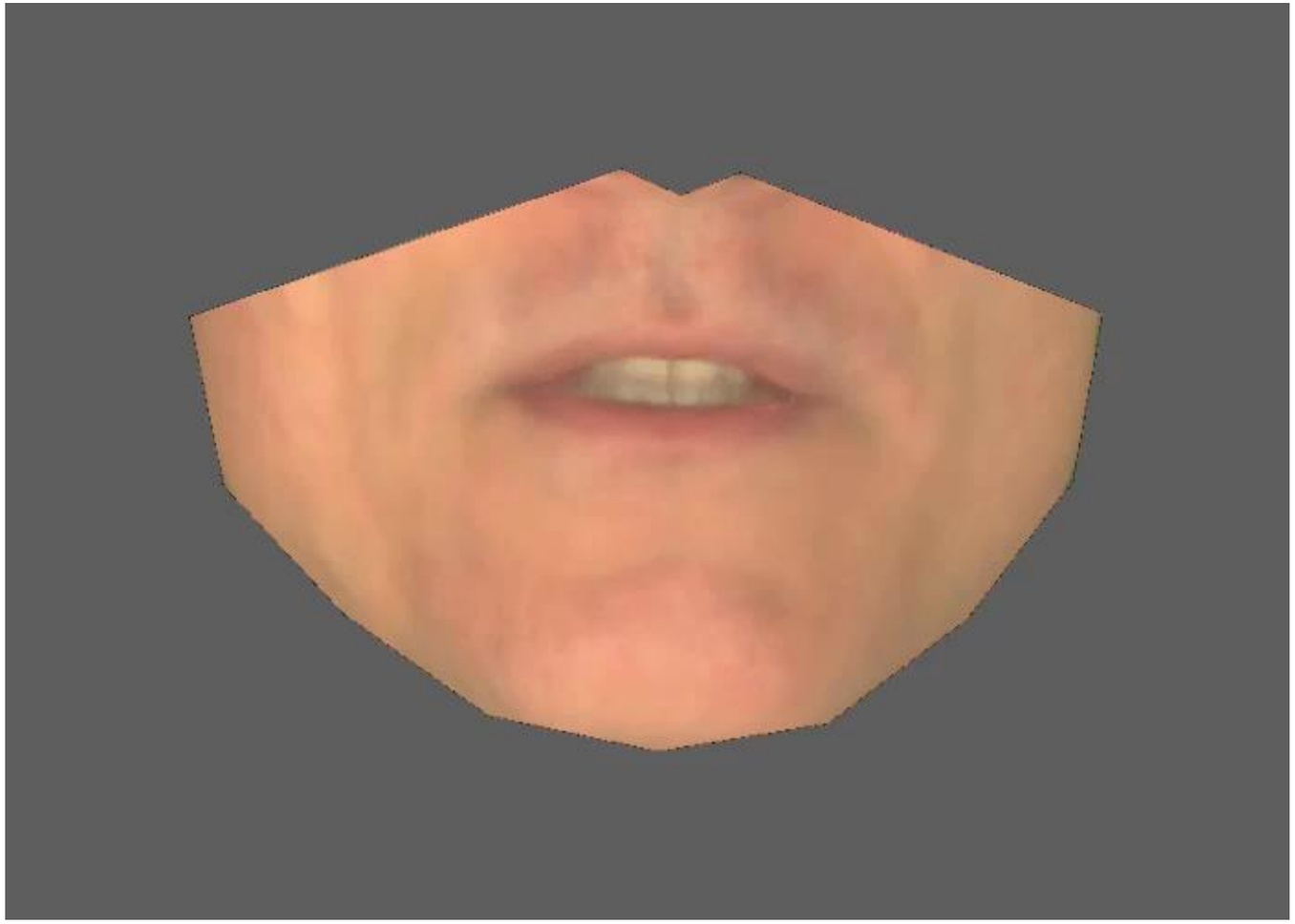
主要内容

- 视位（Viseme）的概念
- 基于参数控制的可视语音合成
  - 视位的参数化描述
  - 视位参数的估计
  - 基于视位的动画生成：
    由静态视位（Static Viseme）到动态视位（Dynamic Viseme）模型
- 基于数据驱动的可视语音合成
  - 视位的视频序列表征
  - 上下文对视位的影响：
    协同发音现象（Co-Articulation）
  - 基于单元选择与拼接的动画生成
  - 图像的平滑过渡变形与叠加合成
- 基于深度学习技术的可视语音合成
  - Demos

Automatically Animate to Input Audio?

A Decision Tree Framework for Spatiotemporal Sequence Prediction
Taehwan Kim, Yisong Yue, Sarah Taylor, Iain Matthews. KDD 2015
A Deep Learning Approach for Generalized Speech Animation
Sarah Taylor, Taehwan Kim, Yisong Yue, et al. SIGGRAPH 2017

Prediction for Very Different Language



Behind the Scenes of Pandora - The World of Avatar https://youtu.be/URSOqWtLix4

Q&A

MEDIA JOINT Research Center