基于汉语视频三音素的可视语音合成

赵 晖　　唐朝京
(国防科技大学电子科学与工程学院　长沙　410073)

摘　要：为了合成具有真实感的视频序列，该文提出一种基于汉语视频三音素的可视语音合成方法。根据汉语的发音规律和音素与视素的对应关系，该文提出“视频三音素”的概念。在此基础上，建立隐马尔可夫(HMM)训练与合成模型，在训练过程中使用了视频音频联合特征，并加入了动态特征。在合成过程中，连接视频三音素 HMM 模型形成句子 HMM，并从中提取特征参数，合成可视语音。从主观和客观评估结果来看，合成视频的真实感强，满意度较高。

关键词：可视语音合成；视频三音素；隐马尔可夫模型；联合特征
中图分类号：TP391.42　　**文献标识码**：A　　**文章编号**：1009-5896(2009)12-3010-05

Visual Speech Synthesis Algorithm Based on Chinese Visual Triphone

Zhao Hui　　Tang Chao-jing
(*College of Electronic Science and Engineering*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: In order to synthesize real video sequence, a visual speech synthesis algorithm based on Chinese visual triphone is proposed. According to Chinese pronunciation principle and the relationship between phoneme and viseme, conception of ‘visual triphone’ is presented. Hidden Markov Model(HMM) is established based on visual triphones. In the training stage, combined features including visual features and audio features are used. In the synthesis stage, sentence HMM is constructed by concatenating triphone HMMs, from which the feature parameters are extracted. From the result of subjective and objective evaluation, the synthesized video is real and satisfied.

Key words: Visual speech synthesis; Visual triphone; Hidden Markov Model(HMM); Combined features

1　引言

可视语音是指人们在用语言交流时所表达出的面部动作，它能在一定程度上传达人们想要表达的意思，帮助人们加深对语言的理解。研究表明，在环境噪声较大或听者有听力障碍的情况下，如果在给出声音信息的同时能给出一个“讲话的头”，则会大大改善人们对声音的理解[1,2]。近年来，在可视语音合成的研究领域取得了较多的研究成果[3–7]，但其所使用的视素模型多是静态的，难以体现视频图像连续变化的特点，且很可能产生跳变现象。

文本针对汉语发音特点，提出视频三音素的分类方法，并在此基础上提出一种基于视频三音素的 HMM 训练和合成模型。

2　汉语语音三音素结构

汉语普通话由音节连接而成，音节由更小的语音单元音素构成，即汉语的声母和韵母。虽然音素可以作为描述汉语普通话的最小单位，但音素在连续语流中很难以稳定形式存在。它受左、右相邻音素的影响，同时又会影响相邻音素，导致其在声学上的表现形式和孤立音节有很大区别。而三音素描写的正是一个音素的稳定段及向左右两边音素的过渡部分，体现了语言的协同发音现象。

一般来说，三音素将声母和韵母作为中心建模单位，并考虑左右音素的影响。三音素模型可以写成 X-Y-Z 的形式， X 代表左面与其相邻的声母或韵尾，Y 代表声母或韵母，Z 代表右面与其相邻的声母或韵头[8]，见表 1。

表 1 三音素的组成

X	Y		Z
	声母	韵母	
a, o, e, er, i, i1, i2, u, ü, -n, -ng, 静音, 21 个声母	b, p, m, f, d, t, n, l, g, k, h, j, q, x, z, c, s, zh, ch, sh, r	a, o, e, i, i1, i2, u, ü, er, ai, ao, an, ang, ia, iao, ian, iang, ua, uai, uan, uang, üan, ou, ong, uo, iou, iong, ei, en, eng, ie, in, ing, un, ün, üe, ui	a, o, e, er, i, i1, i2, u,ü, 静音, 21 个声母

2008-12-05 收到，2009-06-19 改回
国家部委基金(51329060101)资助课题

表 1 中 i1 对应 zi,ci,si 中的 i，i2 对应 zhi, chi, shi 中的 i；而 er 虽然是由单韵母 e 和 r 组成，但其发音比较特殊，在发音过程中口型几乎没有发生变化，仍认为它是单韵母。

3　汉语视频三音素

视素(Viseme)是指与音素(phoneme)相对应的唇部状态。一些针对汉语视素的研究文章[3,4]仅仅考虑了声母和韵母单独发音时所对应的静态视素，由于协同发音现象的存在，利用静态视素合成可视语音时，势必会出现唇部图像不连续或跳跃的情况。为解决此问题，本文根据音素和视素的多对一关系，将汉语三音素精简归类，得到“视频三音素”。采用已有的研究成果[9]，根据唇部特征参数，利用模糊 C-均值算法对音素聚类。将声母聚类为 B，D，J，Z， ZH 共 5 类，将韵母聚类为 A，O，E，I，U 共 5 类，每一类对应一个视素，对表 1 中 X，Y 和 Z 对应的内容归类，见表 2。

根据表 2 的归类结果，简化表 1 中的内容，得到视频三音素的组成，见表 3。

视频三音素在不包括静音的情况下有 625 个，在包括静音的情况下有 750 个，极大精简了表 1 中三音素的类型。为了分析汉语视频三音素的统计特性，需要计算它在汉语语料中的分布概率。从理论上讲，如果某一视频三音素在汉语全部语料中的个

表 2 根据唇部特征对声母和韵母归类的结果

声母类		B	b, p, m, f
		D	d, t, n, l, g, k, h
		J	j, q, x
		Z	z, c, s
		ZH	zh, ch, sh, r
韵母类	X中韵尾分类	A	a, an, ang
		O	ao, o, ong
		E	e, er, en, eng
		I	i, i1, i2, ai, ei, in, ing
		U	u, ü, ou, un, ün
	Y中韵母分类	A	a, ai, ao, an, ang, ia, iao, ian, iang, ua, uai, uan, uang, üan
		O	o, ou, ong, uo, iou, iong
		E	e, er, ei, en, eng, ie
		I	i, i1, i2, in, ing
		U	u, ü, un, ün, üe, ui
	Z中韵头分类	A	a, ai, ao, an, ang
		O	o, ou, ong
		E	e, er, ei, en, eng,
		I	i, i1, i2, in, ing
		U	u, ü, un, ün

表 3 视频三音素的组成

X	Y		Z
	声母	韵母	
A, O, E, I, U, B, D, J, Z, ZH, 静音	B, D, J, Z, ZH	A, O, E, I, U	A, O, E, I, U, B, D, J, Z, ZH, 静音

数为 N_{tri}，而汉语中全部语料所包含的视频三音素的总个数为 N_{total}，则该视频三音素的分布概率

$$P_{\text{dis}} = N_{\text{tri}} / N_{\text{total}} \tag{1}$$

实际情况下，不可能通过计算全部语料来得到 P_{dis}。因此，我们建立了汉语双模态语料库 Bi-VSSDatabase，其中的双模态语料包含视频和音频信息，以视频三音素作为描述连续语音和视频的基本单位和语言现象，能够覆盖各种语言现象，可用来模拟汉语语料的分布情况。根据语料库内容计算视频三音素的分布概率 $P'_{\text{dis}} = N'_{\text{tri}} / N'_{\text{total}}$。当某个视频三音素的 $P'_{\text{dis}} < 1\%$ 时，说明其出现概率很小，不会影响可视语音合成质量。因此，在训练时舍弃此类视频三音素。经统计，符合条件可参加训练的视频三音素的数量为 329 个。

4　特征提取

4.1　唇部特征参数

利用 2 维阈值快速分割算法分割视频图像，提取出唇部特征，如图 1，图中的 7 个参数为原始唇部参数，x 为唇部中线到边缘的距离，u_0 为唇部中线上半部的高度，d_0 为中线下半部的高度，u_1，d_1，u_2，d_2 分别为 x 三分点处相应的高度。在时刻 t 的图像帧，原始唇部参数为 $v_t = [x, u_0, d_0, u_1, d_1, u_2, d_2]$。

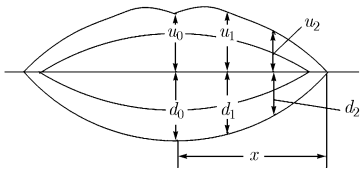


图 1 唇部参数

为了体现唇部运动时的动态特性[5]，计算唇部的动态参数：

$$\Delta v_t = \sum_{\mu=-L}^{L} w_v(\mu) \cdot v_{t+\mu} \tag{2}$$

$w_v(\mu)$ 为权值参数，由原始唇部参数 v_t 和动态唇部参数 Δv_t 组成了 14 维唇部特征参数 $F_{vt} = [v_t, \Delta v_t]$。

4.2　音频特征参数

Mel 倒谱系数(MFCC)符合临界频率和人耳听觉特性，区分能力较强。计算公式如下：

$$C(t,i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg[E_{mel}(t,j)]\cos\left[i(j-0.5)\frac{\pi}{N}\right] \quad (3)$$

N 为三角滤波器个数，$E_{mel}(t,j)$ 为 t 时刻第 j 个滤波器输出的能量，$\{C(t,i)\}_{i=1,2,\cdots,P}$ 为 t 时刻对应的 MFCC 参数，P 为阶数。

同样，为了体现音频特征参数的动态特性，计算语音的动态参数：

$$\Delta a_t=\sum_{\mu=-L}^{L}w_a(\mu)\cdot a_{t+\mu} \quad (4)$$

$w_a(\mu)$ 为权值参数，由原始音频参数 a_t 和动态音频参数 Δa_t 组成了 36 维音频特征参数 $F_{at}=[a_t,\Delta a_t]$。

5 HMM 的训练与可视语音合成

利用双模态语音特征，提出基于视频三音素 HMM 的可视语音合成方法。图 2 为训练过程，从双模态语料库中提取语料并切分三音素，根据视频三音素分类原则对语料归类，分别提取唇部特征和语音特征，形成视频音频联合特征，并将其作为观察矢量训练 HMM 模型，形成视频三音素训练模型集 CVT-HMM$^{(i)}$，$i=1,2,\cdots,329$，模型采用无跨越从左向右模型，状态数为 6。

以句子：“北京欢迎你”为例，拼音为“sil-beijinghuanyingni-sil”，“sil”代表静音，所包含的三音素为 b(sil,ei)，ei(b,j)，j(ei,ing)，ing(j,h)，h(ing,uan)，uan(h,ing)，ing(uan,n)，n(ing,i)，i(n,sil)，根据表 2，得到视频三音素：B(sil,E)，E(B,J)，J(I,I)，I(J,D)，D(I,U)，A(D,I)，I(A,D)，D(I,I)，I(D,sil)。

合成过程如图 3 所示，输入待合成的文本内容，根据三音素切分结果，连缀三音素 HMM 模型，形成对应输入文本的句子 HMM 模型。根据 Viterbi 算法得到文本的联合特征参数 $F_1\ F_2\ \cdots F_T$，采用基于最大似然估计的唇形生成算法得到唇动视频。

6 实验结果

从 Bi-VSSDatabase 选出 31070 个句子。录制时使用高清数码摄像机和麦克风，视频帧速率为 24 Hz，每帧图像的分辨率为 640×360，语音采样率为 44100 Hz，量化值为 16 bit。经统计，预料中包含共 705 个视频三音素，对视频三音素的覆盖率为 94.0%，包括了符合训练条件的全部 329 个视频三音素，能够保证 HMM 的正常训练与合成。

挑选 5 个人的语料(3 男 2 女)共 1000 个句子，每人 200 句，其中 150×5 句为训练用数据，其他 50×5 句用来验证合成效果。当输入“联合国”和“郑和下西洋”两个短句时，输出的视频合成效果如图 4 所示，可以看出，合成的视频图像平稳、连续，真实感强。

图 5 为这两个短句的实际唇部高度(图 1 中 $l=u_0+d_0$)与合成唇部高度的对比曲线。实线为唇部发音的真实高度曲线；点虚线为采用视频单音素的方法合成的视频唇部高度曲线；长虚线为本文方法，即视频三音素方法合成的视频唇部高度曲线。从图中可见，基于单音素的方法存在视频帧跳变的情况，与实际误差较大；而基于视频三音素的方法由于考虑了视频上下文的连贯性，与基于单因子的方法相比，唇部曲线更加平滑，合成的唇部动作接近实际，与实际唇部高度曲线十分吻合。

为了能更好地描述合成结果，需要对输出的合成唇部视频进行主观评价和客观评价。参考 MOS

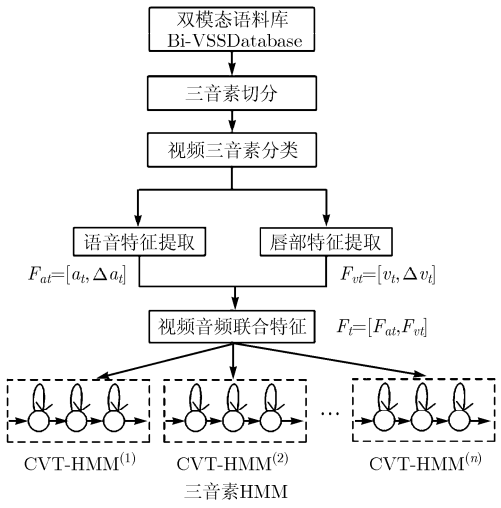


图 2 HMM 的训练过程

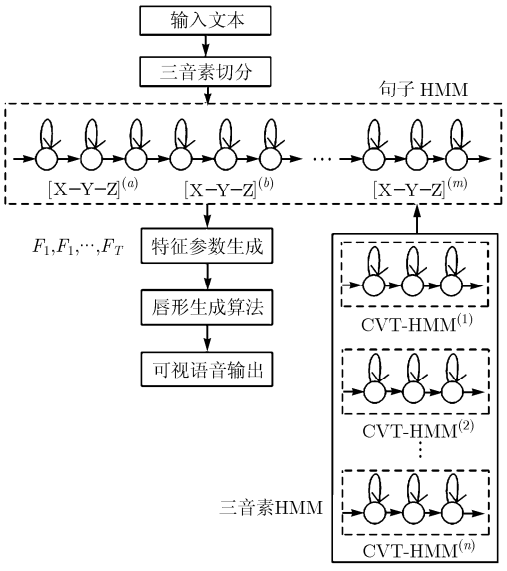
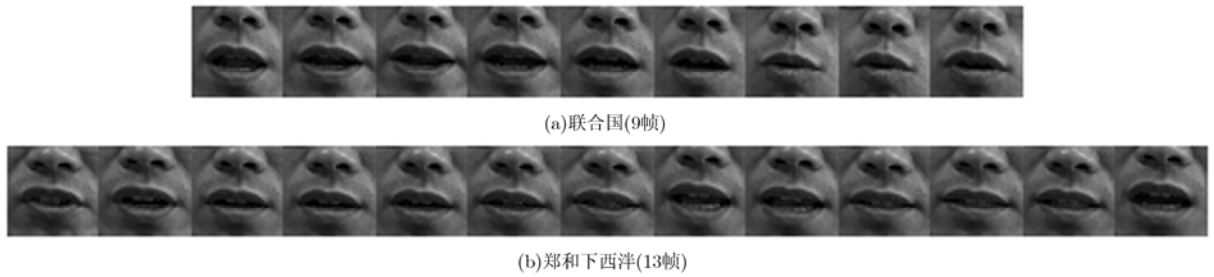


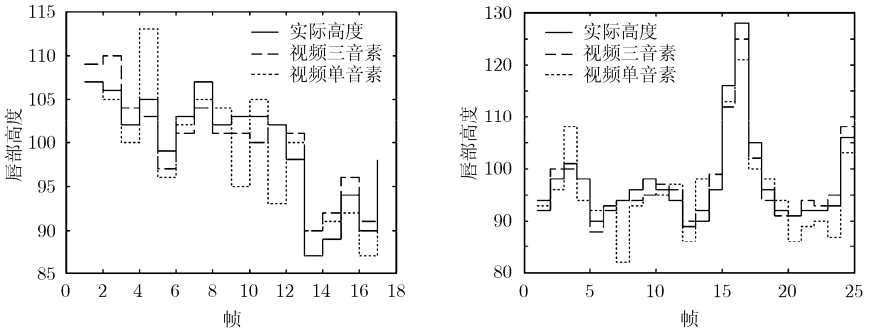
图 3 基于汉语视频三音素 HMM 的可视语音合成过程



(a)联合国(9帧)

(b)郑和下西洋(13帧)

图 4 可视语音合成效果图



(a)联合国　　(b)郑和下西洋

图 5 唇部高度对比曲线

评分方法，定义主观评测标准为 7 分制：5-非常自然，4.5-自然，4-比较自然，3.5-不太自然，3-可接受，2-比较差，1-不能接受。客观评测值也采用同样的 7 分制。

设视频中实际唇部高度序列为 l_{real} (图 5 中的实线)，合成唇部的高度为 l_{syn} (图 5 中的长虚线)。实际高度与合成高度的差值序列为

$$l_{\text{Diff}} = \text{abs}(l_{\text{real}} - l_{\text{syn}}) \tag{5}$$

在对合成视频序列进行客观评测时，既要考虑实际高度序列与合成高度序列的总体平均差距，又要考虑单帧的个体差距，根据 l_{Diff} 的均值 E 和方差 Var 定义客观评测标准，见表 4。

随机抽取 15 个短句并进行主观评测和客观评测，评测值见表 5。

表 4 客观评测值标准

客观评测分	标准定义
5	E<0.1 且 Var <0.001
4.5	E <0.2 且 Var <0.002
4	E <0.3 且 Var <0.004
3.5	E <0.4 且 Var <0.01
3	E <0.5 且 Var <0.03
2	E <0.6 且 Var <0.1
1	$E \geq 0.5$ 且 Var ≥ 0.1

表 5 15 个短句主观评测值和客观评测值比较

短句内容	基于视频单音素方法		本文方法	
	主观评测值	客观评测值	主观评测值	客观评测值
联合国	3	4	4	4.5
郑和下西洋	3.5	4	4	4
国防科技	3	3	3	3.5
好看的电影	4	4.5	4.5	5
永远的画面	3	4	4	4.5
天气很凉	3	3.5	4	4
没完没了	3	3.5	4.5	4
社会主义	3.5	4	4.5	4
排第一	4	4.5	5	5
大江东去	3.5	3.5	3.5	3.5
这座山很高	3.5	4.5	4	5
多多益善	3	4	3.5	4
莎士比亚	4	4	4	4.5
人民解放军	2	3	3	3.5
五星级饭店	2	3	3	3

从主观评测和客观评测的结果来看，用本文方法所合成的可视语音满意度较高，可信度较高且真实感强。而基于单音素的方法中甚至出现了主观满意度比较差的情况(“人民解放军”和“五星级饭

店”)。另外，无论何种方法，客观评测值一般要高于主观评测值，说明人眼对视频中的合成痕迹较为敏感，对真实感的要求较高。

7 结束语

为了实现高质量的可视语音合成，本文提出一种基于汉语视频三音素的可视语音合成方法。实验结果表明，本文方法的合成结果质量较高，图像连续平稳，真实感强，不存在图像跳变的情况。下一步的工作要进一步优化现有 HMM 的模型结构，提高运算速度，以满足实时性要求。

参 考 文 献

[1] Summerfield Q. Use of visual information in phonetic perception[J]. *Phonetic*, 1979, 36(4/5): 314–331.

[2] McGurk H and Macdonald J. Hearing lips and seeing voices[J]. *Nature*, 1976, 264(5588): 746–748.

[3] Perng Woei-luen, Wu Yung-kang, and Ming Ouh-young. Image talk: a real time synthetic talking head using one single image with Chinese text-to-speech capability[C]. Sixth Pacific Conference on Computer Graphics and Applications, Singapore, 1998: 140–148.

[4] 王志明, 蔡莲红, 吴志勇. 汉语文本-可视语音转换的研究[J]. 小型微型计算机系统, 2002, 23(4): 474–477.
Wang Zhi-ming, Cai Lian-hong, and Wu Zhi-yong. Study of text to visual speech in Chinese[J]. *Mini-Micro-System*, 2002, 23(4): 474–477.

[5] Masuko T, Kobayashi T, and Tamura M, *et al*.. Text-to-visual speech synthesis based on parameter generation from HMM[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, USA, 1998, 6: 3745–3748.

[6] Jiang Jin-tao, Aronoff J M, and Bernstein L E. Development of a visual speech synthesizer via second-order isomorphism[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, USA, 2008: 4677–4680.

[7] Zhou Wei and Wang Zeng-fu. Speech animation based on Chinese mandarin triphone model. 6th IEEE/ACIS International Conference on Computer and Information Science, Melbourne, Australia, July 2007: 924–929.

[8] 吴华, 徐波, 黄泰翼. 基于三音素模型的语料自动选取算法[J]. 软件学报, 2000, 11(2): 271–276.
Wu Hua, Xu Bo, and Huang Tai-yi. Automatic corpus selecting algorithm based on triphone models[J]. *Journal of Software*, 2000, 11(2): 271–276.

[9] Zhao Hui and Tang Chao-jing. Visual speech synthesis based on Chinese dynamic visemes[C]. IEEE International Conference on Information and Automation, Zhangjiajie, China, June, 2008: 139–143.

赵 晖： 男，1980 年生，博士生，研究方向为多媒体通信、可视语音合成和网络图像安全等.

唐朝京： 男，1962 年生，教授，研究方向为多媒体通信、网络攻防对抗等.