语音同步的可视语音合成技术研究

贾熹滨，尹宝才，李敬华
（北京工业大学 计算机学院 多媒体与智能软件技术北京市重点实验室，北京 100022）

摘 要：为了提出一种真实感较强的可视语音合成方案，对目前国内外主流研究方法进行了探讨．在基于对可视语音合成问题分析的基础上，提出了可视语音合成系统研究方法中首先要解决的2个问题：视觉语音特征模型的构建和声视频映射模型的构建．分析了目前国内外研究方法的主要解决方案，提出了在未来研究中本系统将采用的系统框架和重点研究内容．

关键词：语音动画；声视频映射；特征定位；人脸建模
中图分类号：TP 391 **文献标识码**：A **文章编号**：0254－0037(2005)06－0656－06

具有真实感的语音同步的可视语音合成技术的研究是近年来人机交互领域一个重要的研究方向，将具有语音动画的人头，如桌面代理、虚拟播音员、虚拟销售员，用在教学、通信、电子商务、广播、影视等领域可进一步提高人机交互的友好性和方便性．可视语音合成系统设计过程中首先面临的问题是语音和可视语音相关性证明以及声视频关联关系的构建．对于声视频之间的相关性，早在20世纪70年代Q. Summerfield等就从人类理解语言时声视频所起作用的角度出发做了一些测试，通过实验证明：在有噪背景下，与只提供声音信息相比，同时提供声音和同步的说话人脸视频图像，实验对象对语言的理解正确率提高31%[1]．Massaro和Cohen等通过不同实验得出了类似的结论，证明测试者在声音语音、视觉语音(唇动)以及双模式(声视频语音)3种不同发音环境下对语言识别的正确率分别为55%、4%、72%[2]．另一些研究者提出的称为"McGurk effect"的现象[3]，从另一个角度验证了声音和口形运动的一致性对人类理解语言的影响，在一定程度上证明了声视频在描述语言时所提供的信息是相关的、一致的．但上述实验仍不足以完全说明声视频之间的具体关联关系，对于声视频关联关系模型的构建目前研究方法还没有统一的定论，主要还是基于具体假设条件下，通过建立一种合理的声视频模型加以描述．由于人脸是一个相对复杂的物体，加上说话时人脸运动的复杂性和随机性，如何描述视觉语音的静态模式及动态模式成为整个系统设计过程中的一个重点及难点．基于以上分析可以看出，可视语音合成系统研究方法中首先要解决2个问题：①视觉语音特征模型的构建；②声视频映射模型的构建．

1 视觉语音特征模型

语音同步的人脸特征模型的建模实质上需要解决2个问题：人脸静态建模和语音动画处理．目前国内外主要的可视语音合成系统所采用的解决方案可分为2类：一类是借鉴人脸三维建模和动画处理技术的研究成果，采用基于模型的方法，通过建立语音和人脸模型之间的对应关系实现可视语音动画的合成；另外一类方法则直接利用采集的说话人脸图像来表示人脸，通过分析语音的动态信息预测出对应的口形运动，并通过诸如轨迹处理技术、变形技术(morphing)等图像处理技术来合成语音同步的语音动画．2类方法各有优缺点，不同系统也有各自不同的解决方案．

1.1 基于图形学的处理方法

基于图形学的处理方法，又称为基于模型的方法、基于知识的方法．其设计思路是构建一个三维的人

收稿日期：2004-05-12．
基金项目：国家自然科学基金资助项目(60375007)．
作者简介：贾熹滨 (1969－)，女，山西太原人，讲师．

脸参数模型，通过学习说话过程中人脸参数点的运动规律，建立语音信号和语音动画参数模型之间的映射关系．在合成阶段根据对新输入语音信号的分析和已有的先验知识控制形状参数的变化，生成与语音同步的口形/人脸动画．

根据模型参数建立机制不同，基于模型的方法可以分为 2 类：一类是从生理解剖学的角度，以组织学、生物力学、运动学和面部骨骼结构知识为基础，利用计算机图形学的方法构建皮肤、肌肉、骨骼等部分的三维模型，并将某一特定语音信息映射为特定的参数状态，如美国加州大学感知科学研究室(PSL-UCSC)的 D. Massaro 等人提出直接建立语音音素与发音器官几何位置参数之间的映射关系，并采用传感设备来捕获人发音中参数点运动轨迹值，通过人机交互方式实现语音音素与模型参数对应关系的设置[4]．基于模型的方法中语音动画的仿真目前多采用建立人脸特别是唇部周围肌肉等的弹性模型的方法，利用动力学原理近似肌肉的运动，通过移动肌肉、拉伸皮肤来实现．另一类方法是建立在二维或三维网格几何人脸参数模型基础上．J. P. Lewis 等人提出了一种自动语音同步的唇动合成系统解决方案[5]，该方法采用基于 F. Parke 提出的三维网格几何模型实现对人脸静态模型的建模，在用线性预测算法实现语音音素识别的基础上，通过建立音素与相应参数点之间的映射关系，在合成过程中找到语音同步的关键帧人脸参数点的位置，最后通过轨迹平滑技术获得连续口形动画．陈益强等人提出利用 3 层前馈神经网络建模语音特征和 FAP 模式之间的统计关系，在合成阶段对新输入语音用该模型分析预测出 FAP 模式，并驱动基于 MPEG-4 定义的网格模型，合成出语音同步的语音动画[6]．

利用基于模型方法实现语音动画合成的最大的优点就是灵活性强，可通过控制参数获得所需要的语音动画效果．但由于说话时人的面部特别是唇部的运动是一种非刚性的自发的运动，用物理或网格模型完全再现人说话时唇部周围复杂的肌肉解剖结构、运动规律以及相应皮肤拉伸、骨骼运动等确实是难度很大的工作，需要丰富的先验知识和复杂设计模型的保证，加上说话行为是人类最熟识的行为，因而基于模型的方法在合成效果上总让人感觉有些机械、不自然，合成效果真实感较差．

1.2　基于图像处理的方法

基于图像处理的方法，又称基于样本、基于数据驱动的方法，是近年来一些研究者提出的一种通过样本采集、学习和预测的方式合成语音同步的口形/人脸动画的方法．该方法由于其合成效果真实感强，很快成为可视语音合成研究方法中的一个热点．该方法首先利用数据采集设备获得说话对象的声视频序列，分别对声视频序列进行预处理，提取声视频特征，建立样本数据库；在此基础上，根据声视频映射模型构建的假设条件对样本进行标识和训练，建立声学语音特征和视觉语音特征之间的关联关系；最后在合成阶段利用这种关联关系预测出与新输入语音同步的视觉语音特征序列，并通过样本库搜索和轨迹平滑获得最后的语音同步视觉语音动画序列．可以看出基于图像处理的声视频合成系统在某种程度上说是一种采集图像的重组，因而合成效果相对基于模型的方法更逼真．在基于图像处理的方法中，动态模型描述的是一种声视频映射规则的定义及唇部运动轨迹再现和平滑处理的机制．因而不同于基于参数模型方法，基于图像处理的方法关键在于 2 点：第 1，图像的表示，即图像特征的选择，用来实现训练阶段图像计算及合成阶段图像匹配；第 2，声视频映射机制的描述，用以预测语音同步的视觉语音动画．

视觉语音即说话过程中不同的口形，主要来自说话器官几何形状和纹理的变化，基于像素级的图像表示由于存在大量冗余信息，在实际系统中更多采用基于特征表示的方法．目前方法中比较典型的有：基于全局特征方法、基于局部特征方法以及基于全局和局部特征结合的方法．采用不同特征表示，其系统设计也各不相同．

T. Ezzat 等提出的称为 Mike Talk 的基于文本驱动声视频合成系统中所采用的图像表示方法是基于像素级对应的方法[7]，该系统设计的核心是建立语音音素和关键帧图像对应数据库，其基本模型选择了 16 个具有较大口形差异的视素(viseme)作为关键帧，中间帧则通过光流技术对相邻关键帧之间进行对齐后，利用变形技术合成．但这种方法对说话过程中连读问题较难解决．

C. Bregler 等人提出的称为 VideoRewrite 系统中采用了大样本采集和重组的合成方法[8]，该方法中的静态模型采用描述嘴唇和下颌轮廓的 54 个特征点，口形动画处理采用建立音素和视频图像映射模型，

并利用搜索机制找到语音同步的口形序列. 为了解决连读问题,提出了建立三音素数据库方法,在搜索中采用以中间音素为核心,通过计算音素和相邻唇形之间的加权距离,在已标定音素、口形特征的数据库中找到最相似口形图像,然后缝合到具有不同姿态和习惯动作的人脸上,合成语音同步的可视语音序列. 该系统提出了一种自动特征点定位的方法,只需手工标定26幅图片,在此基础上,利用morphing技术将26幅标定了特征点的图片扩展为351幅具有标定信息的图片,通过对扩展数据集进行学习得出Eigenpoints主特征点模型,建立特征点向量和图像之间的线性映射关系,实现特征点的自动标定. 考虑到建立所有特征点和图像之间的线性映射关系不准确,分别建立了唇部和下颌的Eigenpoints模型.

M. Brand提出的基于双输出HMM模型的可视语音动画合成系统(voice puppetry)中利用26个特征点的位置及速度矢量表示人脸特征[9],动态模型采用连续隐马尔可夫建模人脸运动,通过轨迹运算算法完成动画处理. 该模型可应用在人脸和非人脸以及动画形象的语音动画中,可驱动参数模型又可以通过样本数据库搜索合成最终可视语音动画,灵活性强,能实现对非特定人的处理.

E. Cosatto等人提出了另一种基于大样本的可视语音合成方法[10],该方法中采用唇宽、上下唇高等几何参数,连同整体图像的30维PCA系数一起作为每帧图像的静态模型. 在训练阶段建立音素与对应口形特征数据库;在合成阶段用音素作为搜索条件搜索样本库找出若干对应口形,求出同类口形的几何参数均值,选择具有该几何参数值的若干候选口形,最后根据相邻口形最相似的原则,利用Viterbi算法通过计算相邻口形图像PCA系数间的欧氏距离,求得具有最小欧氏距离的口形序列. 在该方法中,准确的特征定位是影响合成效果的主要因素. 为了获得准确的特征点的位置,采用了由粗到精渐进的特征定位算法,首先利用颜色分割和纹理分割算法实现人脸特征粗定位,在此基础上通过学习人脸的主要特征尺寸和相互位置关系,进一步确定人脸特征. 随后利用事先学习出的局部人脸特征颜色模板,确定各人脸特征的准确形状. 最后对上步中标定出的每个特征利用模板匹配方法确定出特征点的准确位置. 考虑到头部姿态和光照条件的影响,文献[10]提出了利用多个核函数进行相关性分析的特征点定位方法,如嘴角利用9个不同核函数,眼角采用3个不同的核函数. 为了提高运算速度,系统还采用可变梯度计算方法代替全局运算计算特征相关性.

与基于参数的可视语音合成方法相比,基于数据驱动的合成方法是一种样本学习和预测重构的过程,合成效果真实感更强,模型复杂性也相对较低,且不需要过多的口形运动先验知识. 该方法的缺点是需要较大的数据存储空间,另外特征点的准确定位也是目前难度较大的问题. 如何更充分地利用说话过程中口形运动的形状信息及纹理信息等局部特征来解决视觉语音表示问题是目前方法的一个研究点.

2 声视频映射模型

目前国内外所提出的声视频映射机制总的来说可以分为两大类,一类是假设语音音素(phoneme)和对应口形视素(viseme)之间存在确定的对应关系;另一类是基于语音声学特征建立声视频映射模型.

基于确定对应关系的声视频映射借助了语音识别的成果,利用语音识别技术将语音序列识别为音素,然后建立音素与视素之间的映射关系[5,8]. 这类基于语义层的音素和视觉语音映射关系,易于实现,并易于根据对语义的理解在映射关系上建立一些处理机制,如VideoRewrite系统采用了记录三音节方法,将上下文信息包括进来以处理连读对实际口形的影响[8]. 该方法的缺点是:对语音识别技术有很大的依赖性;识别为音素后会失去影响合成结果逼真性的韵律特征;忽略了音素上下文相关性的影响.

考虑到语音合成的目的在于预测口形而不是语义理解,近年来一些研究人员提出了一种直接建立语音声学特征和唇动之间映射关系的解决方案[8-12]. 建立基于声学层的声视频映射关系,可充分利用语音的信息挖掘口形的运动规律. 但由于没有语义信息,其对应的准确性和所建模型有很大关系.

2.1 基于矢量量化的映射机制

矢量量化方法的基本原则是建立2个对应的码本[11]:语音码本和视觉语音码本. 该方法首先对训练集中样本的语音特征进行聚类形成语音码本,然后将对应的可视语音映射到具有相同码字的可视语音码

本中．在合成阶段，利用所建码本，将新的语音特征与语音码本进行比较，按照距离最小原则分类到相应类，输出对应码字，将该码字所对应的可视语音码本中的均值图像作为输出．该方法简单易行，且可以通过加大码本长度来提高精度，但量化误差带来了合成图像的跳跃问题．另外该方法对视频序列的预测完全依赖于语音特征的聚类特性，选择能够反映口形变化的语音特征就显得非常重要．

2.2　基于神经网络的映射模型

基于神经网络的映射模型是用神经网络学习方法，通过样本学习建立语音特征空间到视觉语音特征空间的映射模型．根据所采用的神经网络模型不同，该方法主要包括基于多层反馈神经网络模型和基于时延神经网络模型的 2 种方法．陈益强等所提的方法[6]，采用了 3 层反馈神经网络结构，利用包含前后 6 帧的 18 维语音特征向量作为输入节点．F. Lavagetto 提出的系统中采用了时延神经网络(TDNN)模型[12]，直接从模型结构反映语音的时序关系，在一定程度上解决了连读问题．

基于神经网络的方法，在合成时计算效率高，实时性较好，为语音和可视语音关联关系模型化提供了一种有效的解决方案．其缺点是：①神经网络结构(如隐藏层的数目)的确定通常是基于经验来定义的，缺乏理论支持．②目前的模型还无法解决自然语音上下文相关信息的不定长问题，因此影响了连读发音的仿真．③基于神经网络建立的声视频映射模型描述的是语音特征和视觉语音特征之间的一种确定性映射关系，无法完全反映唇动的复杂性和随机性．为解决上述问题，一些研究人员利用概率统计模型来建立声视频的映射关系，主要包括 HMM、高斯模型、动态贝叶斯网以及支持向量机．鉴于 HMM 在语音识别领域的成功应用和 HMM 在处理时序问题上的优点，目前的系统多采用基于 HMM 的方法．

2.3　基于 HMM 的映射模型

基于 HMM 的映射模型通过学习模型参数建立语音和可视语音信息之间的概率统计关系．合成时对新输入的语音序列，利用训练阶段建立的模型，按照最大概率密度、最小均方差等原则找到可视语音参数序列．将 HMM 用在可视语音合成系统中，设计的重点在于如何用 HMM 的状态和各状态对应的输出分布建立起反映语音和可视语音关系的模型．目前的研究方法提出了多个不同的设计机制，其设计的机理是建立在对描述语音和可视语音状态转移含义的不同理解基础上，并且提出了不同的 HMM 设计模型．

由 Yamamoto 等人提出的方法的设计机理是基于声学语音音素 HMMs 模型[13]．对于每个音素 HMM 模型，其状态对应的是视觉语音．核心步骤是：在训练阶段，用语音作为观察值去训练，然后用 Viterbi 算法求出语音的最佳状态序列，并将状态与对应口形对齐，对训练例中每个状态对应的口形参数求解平均值，建立包括音素、口形参数平均值的数据库；在合成阶段，对新的输入语音用 Viterbi 算法求出最佳状态序列，然后找出对应状态的口形参数．考虑到连读问题，该方法中还提出了训练双音节、三音节的隐马尔可夫模型的解决方案．由于该方法语音采用了连续高斯模型，克服了声学特征量化的误差，相对矢量量化的方法平均误差降低了 8.7%～32%[13]．在随后的工作中，Yamamoto 等人对该方法进行了进一步改进，提出了利用期望值最大算法(EM)代替 Viterbi 算法，将估计误差降低了 26%．由于模型训练只利用了语音信息，易受噪声影响，另一方面没有充分考虑视觉语音信息，另外由于状态与视觉语音平均值对应，因而所合成的视觉语音仍存在量化误差．

为使隐马尔可夫模型既涵盖语音又涵盖视觉语音，一些研究人员提出了用语音和可视语音联合训练隐马尔可夫模型的方法．如 T. Masatsune 等人所提出的基于参数模型的语音-文本驱动唇动合成系统[14]，在训练阶段，用语音特征和视频特征一起作为观察值，训练音节 HMMs，每个音节 HMM 的输出包含语音特征和视觉语音特征 2 部分；在合成阶段利用训练阶段获得的隐马尔可夫模型，对新输入语音进行处理，识别为若干音节 HMMs 序列，再根据对应的 HMM 计算出语音同步的视觉语音序列．该方法中是建立在同一音节 HMM 的语音和可视语音输出具有相同的观察概率密度的假设上．

R.R. Rao 等人提出了另一种基于联合训练的方法[15]，直接建立语音声学层特征和可视语音特征之间的关联关系．该方法采用的机制是建立联合隐马尔可夫模型中状态与语音和可视语音联合混合高斯概率分布模型之间的对应关系．在训练阶段，使用声视频特征作为观察向量进行模型训练，随后通过对视觉

参数积分从混合模型的联合观察概率密度中求得语音 HMM 模型的观察概率密度,并继承其他参数建立语音 HMM;在合成阶段,基于训练阶段所得的语音 HMM,对新输入语音利用 Viterbi 算法求出最佳状态序列,找出每个状态对应的语音和可视语音联合混合高斯概率分布,然后依据最小均方差准则估计出连续的可视语音序列. Choi 等人提出的 HMMI 方法[16]采用了类似的机理,提出了基于 Baum-Welch 的重估方法,依据最大概率密度估计准则推出可视语音序列. 该方法相对前一种方法可避免将输入语音的噪声直接传到输出的估计中,具有更准确的预测结果.

不同于建立基于语音单位的多个隐马尔可夫模型,M. Brand 提出的双输出隐马尔可夫模型采用了一种全局 HMM 模型结构[9],其状态对应的是在说话时具有较大差异的可视语音. 声视频映射机理建立在语音和可视语音具有相同模型结构的假设基础上. 在训练阶段,以视频信号作为观察值进行模型训练,然后通过反映射,得到具有双输出的隐马尔可夫模型;在合成阶段利用 viterbi 算法对新输入语音求得一个最佳状态序列,从而找出对应的可视语音参数,并利用轨迹运算算法求得平滑的可视语音序列. 该文献提出了基于熵最小原则的隐马尔可夫模型结构估计和训练算法,为描述可视语音之间的转移关系提供了合理的模型,降低了计算复杂度.

J. Williams 等人将声视频频率不一致的特性考虑进去,提出了一种新的映射机制[17],通过建立关联 HMM 模型来描述语音和视觉语音之间的关联关系,具体结构设计为:关联 HMM 的状态转移矩阵及初始分布继承视觉 HMM 模型的值,观察概率密度延续下采样处理后语音特征的输出概率. 该方法克服了联合训练带来的模型与特定人相关的弱点,提高了模型的通用性. 总之,基于 HMM 的方法不失为解决可视语音合成问题的一个有效方法,而新的系统的设计必须从解决模型对实际语音和口形动画仿真入手,借鉴上述系统映射关系构建机制提出更合理的解决方案.

3 结束语

作者围绕解决可视语音合成技术的 2 个关键问题对目前国内外的一些方法进行了探讨,分析了目前在解决视觉语音建模和声视频映射模型问题中所采用的设计机制和关键技术,特别对各方案的合成效果以及在具体实现时的优缺点作了分析. 基于以上理解和分析,在下一步的工作中,围绕如何提高合成效果的真实感,提出了以下设计思路和研究方向.

综合分析上述方法,在视觉语音特征建模问题上,基于图像的方法在真实感处理上有相对明显的优势,不需要获取过多的先验知识并且可以获得很好的合成效果. 考虑到上述原因本系统欲采用基于图像处理的方法来建模视觉语音. 在具体的技术设计和实现中可借鉴基于参数模型方法中的一些经验:利用基于参数模型方法中已有的一些知识作为约束条件提高模型学习的收敛性;在选择图像特征模型时,可选择在合成端兼容搜索重建方式和参数模型驱动方式的特征模型.

在声视频映射模型构建方面,利用隐马尔可夫模型建立语音和可视语音的关联关系有以下优势:第 1,能有效处理连读问题;第 2,能提供灵活的模型设计结构,完成对语音和视觉语音映射关系的模拟;第 3,具有较成熟的求解算法的支持. 基于以上原因,本系统欲采用隐马尔可夫模型对语音和可视语音之间的关联关系建模. 在具体实现方案上,借鉴以上方法的经验,并综合对人说话时生理习惯的理解,提出合理的声视频映射机制.

目前在语音特征提取方面还没有过多的探讨,而如何提取与语音动画相关的语音特征模型是提高整个系统合成效果的关键点之一. 在未来研究工作中,可以考虑利用目前语音研究领域的一些成果,对主要的语音特征(包括时域特征、频域特征及韵律特征)对预测可视语音序列的贡献进行学习,选择与语音动画相关的特征,建立合理的模型.

参考文献:

[1] SUMMERFIELD Q. Use of visual information in phonetic perception[J]. Phonetica, 1979, 36(4-5): 314-331.

[2] COHEN M, MASSARO D. Modeling coarticulation in synthetic visual speech in models and techniques in computer anima-

tion[A]. Computer Animation '93[C]. Tokyo: Springer-Verlag, 1993. 139-156.

[3] MCGURK H, MACDONALD J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746-748.

[4] COHEN M, MASSARO D, CLARK R. Training a talking head[Z]. ICMI'02, IEEE 4th Int Conf on Multimodal Interfaces, Pittsburgh, 2002.

[5] LEWIS J P, PARKE F. Automated lip-synch and speech synthesis for character animation[Z]. CHI/GI 1987 Conference on Human Factors in Computing Systems and Graphics Interface, Toronto, Canada, 1987.

[6] 陈益强，高文，王兆琪，等. 基于机器学习的语音驱动人脸动画方法[J]. 软件学报，2003，14(2)：215-222.
CHEN Yi-qiang, GAO Wen, WANG Zhao-qi, et al. A speech driven face animation system based on machine learning[J]. Journal of Software, 2003, 14(2): 215-222. (in Chinese)

[7] EZZAT T, POGGIO T. Visual speech synthesis by morphing visemes[J]. International Journal of Computer Vision, 2000, 38(1): 45-57.

[8] BREGLER C, COVELL M, SLANLEY M. Video rewrite: Driving visual speech with audio[A]. Proc ACM SIGGRAPH' 97[C]. New York: ACM Press/Addison-Wesley Publishing Co, 1997, 353-360.

[9] BRAND M. Voice puppetry[A]. Proceedings of ACM SIGGRAPH 1999[C]. New York: ACM Press/Addison-Wesley Publishing Co, 1999. 21-28.

[10] COSATTO E, GRAF H P. Photo-realistic talking-heads from image samples[J]. IEEE Transactions on Multimedia, 2000, 2(3): 152-163.

[11] KAKUMANU P, GUTIERREZ OSUNA R, ESPOSITO A, et al. Speech-driven facial animation[A]. Proceedings of the 2001 workshop on Perceptive user interfaces[C]. Orlando: ACM Press, 2001. 1-5.

[12] LAVAGETTO F. Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio-video synchronization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1997, 7(5): 786-801.

[13] YAMAMOTO E, NAKAMURA S, SHIKANO K. Lip movement synthesis from speech based on hidden markov models[J]. Speech Communication, 1998, 26(1-2): 105-115.

[14] MASATSUNE T, SHIGEKAZU T M, TAKAO K. Text-to-audio-visual speech synthesis based on parameter generation from HMM[Z]. Sixth European Conference on Speech Communication and Technology, Budapest, 1999.

[15] RAO R R, CHEN T. Audio-to-visual conversion for multimedia communication[J]. IEEE Transactions on Industrial Electronics, 1998, 45(1): 15-22.

[16] CHOI K, LUO Y, HWANG J N. Hidden markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system[J]. Journal of VLSI Signal Processing, 2001, 29(1-2): 51-61.

[17] WILLIAMS J, KATSAGGELOS K. An HMM-based speech-to-video synthesizer[J]. IEEE Transactions on Neural Networks, 2002, 13(4): 900-915.

A Survey on Speech-synch Visual Speech Synthesizing Techniques

JIA Xi-bin, YIN Bao-cai, LI Jing-hua

(Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing 100022, China)

Abstract: In order to get a kind of feasible designing scheme to improve the realistic effects, the main research method is discussed. Based on that, two key questions are proposed: One of them is constructing the visual speech representation model, the other, audio/visual mapping model. After analyzing the main resolution scheme both nationally and internationally, system scheme and key research contents are proposed in the end.

Key words: visual speech animation; audio/visual mapping; feature location; face modeling