基于 B/S 模式的 3D 双语虚拟说话人的研究与实现

林会杰　贾珈　王晓慧 蔡莲红

普适计算教育部重点实验室

清华信息科学与技术国家实验室（筹）

清华大学计算机科学与技术系，北京，100084

摘　要：随着计算机技术的不断发展，人们希望计算机具有更多的智能化和人性化，能够模拟人的方式与使用者进行交流。同时，人们也希望程序能够方便的获取、安装和使用。文本到可视语音的转换技术，正是实现计算机人性化的重要组成部分。

本文基于本实验室的 3D 汉语虚拟说话人平台，采用视位映射的方法，建立了英语音素到汉语视位的映射，实现了支持英语和汉语两种语言的 3D 虚拟说话人，为实现统一参数的多语种虚拟说话人系统提供了一种解决方案；建立了一种基于分类处理的协同发音模型，提升了英语虚拟说话人口型的自然度。

进而，本文采用 B/S 模式，通过 ActiveX 技术，将其封装为可以在浏览器中发布运行的插件，用户可以在浏览器中方便的访问、使用虚拟说话人服务，其质量和性能与传统桌面版虚拟说话人系统相同。

关键词：视位映射；协同发音；虚拟说话人；B/S 模式；ActiveX

1．引言

虚拟说话人通过结合语音和视频两种模态的信息，在音视频同步的基础上，模拟人说话过程中的口型、舌头等可视发音器官的变化，为人机交互提供了更为自然和友好的交互方式，在动画制作、交互式游戏设计、多媒体交互、电子课堂等领域中，具有广泛的应用前景。

在以往的虚拟说话人系统的研究中，为了获得准确的静态视位，往往采取从录制的语料库中训练、提取视位参数的方法，首先针对一种特定的语言构建需要的训练语料库，然后采用一定的算法从中提取视位参数[1,2]。采用这种方式能够获得较为准确的静态视位，但对于多语种的虚拟说话人系统，采用这种方式将需要对每一种语言构建语料库、训练提取参数，提取的参数也难以统一。

协同发音问题指的是在连续语流中，一个音素的发音和口型会受到其前后相邻的音素的影响。在连续语流中，协同发音对视位有着明显的影响。在以往的协同发音解决方法中，Cohen 和 Massaro 提出的基于控制函数的模型[3]是较为常用的方式。在该处理模型中，需要根据语料库确定控制函数类型和控制参数。而控制函数类型和控制参数不仅与语言相关，也与具体的说话人有一定的关系，因此，这种从语料库中训练得到的协同发音模型只适用于特定的说话人。

联系作者：林会杰，E-mail：linhuijie@gmail.com

区别于传统的虚拟说话人实现方式，本文提出了不同语言间视位映射的方法，在已有的汉语视位参数模型的基础上，根据英语音素的发音部位和类型，建立了英语音素到汉语视位的映射，为实现多语种的虚拟说话人系统提供了一种可行的解决方案；通过建立基于分类处理的英文协同发音模型，实现在连续语流中具有较高自然度的英汉双语 3D 虚拟说话人，这种方式与具体的说话人无关，适用于可更换人脸模型的情况；利用 ActiveX 技术，将最终的系统封装为浏览器插件，使得用户可以在浏览器中访问、使用该虚拟说话人系统服务，且与桌面版虚拟说话人相比，具有相同的性能和质量。最后，本文设计了一组主观评测实验，证明了采用视位映射的方法实现的英文版虚拟说话人具有很好的效果，证实了视位映射的方法的可行性和有效性。

2．视位映射与协同发音模型

本文提出的利用视位映射的方法实现的 3D 双语虚拟说话人系统框架如图 1 所示：

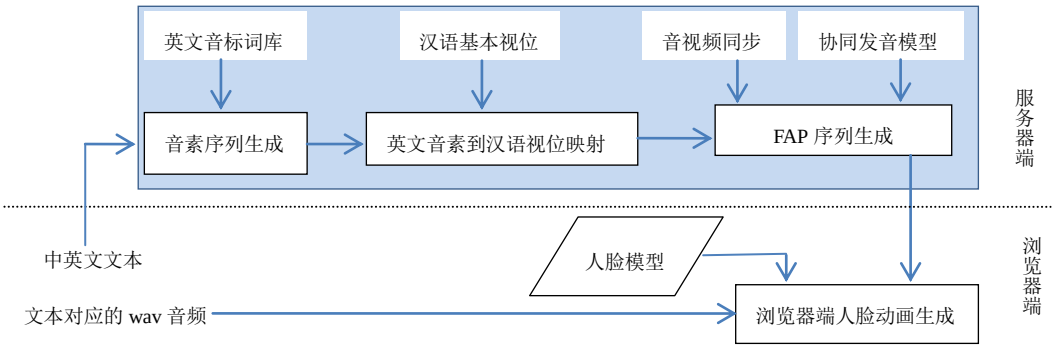

图1 系统整体框架图

在上图中，系统的输入是中英文文本、以及文本对应的 wav 音频文件。客户端与服务器端通过 Socket 进行传输通信。其中虚线上方的模块均是在服务器端，服务器端首先根据输入建立英文单词到音素的转换，获取英文文本的音素序列，然后根据英语音素的发音部位、发音方式等特点进行分类，获取对应的汉语视位表示，汉语视位参数基于本实验室的汉语虚拟说话人 Talking Avatar[2]提取的参数，进而利用英语协同发音规则，改善连续语流中英语发音口型，并通过增加时间标记的方法，实现音频与视频的同步，最终生成相应的 FAP 序列，传送给浏览器客户端，浏览器客户端根据生成的参数，调用底层 OpenGL 库，绘制人脸动画。

2.1 英文单词到音素转换

英文单词到音素（Phoneme）转换，是指将英文单词用其对应的音素序列进行表示，如英文单词“word”对应的音素序列是：\w\，\er\，\d\。英文单词到音素的转换，可以由 TTS 提供，在没有 TTS 的情况下，也可以采用词库查询的方式。

在系统的实现中，本文采用的是基于词库的方法，英文音标词库采用的是 CMU 研发的开源词库 CMUDict v0.7a 版本[4]。CMUDict 包含了超过 125000 个引文单词的北美发音音标，具有较高的准确性和完善性，可以满足虚拟说话人系统的需求。

2.2 英语音素到汉语视位映射

在语音学中，音素（Phoneme）是某一语种的最小可区分的发音单位。与音素相对应的人的发音器官所处的状态，称为视位（Viseme）。音素是与具体的语言相关的，每一种语音都有其对应的音素集合，如国际音标表示中，英语由 48 个音素组成；汉语中有 21 个声母和 38 个韵母组成，可以划分为 32 个音素。

在对比语音学研究[5]中发现，英语与汉语语言间的音素在发音部位、发音方式等方面有很大的相似性，这就使得表示汉语音素口型的视位参数，可以用来表示具有相似发音部位和发音方式的英语音素的口型。比如，英文音素\f\与汉语中声母\f\的发音部位和发音方式几乎相同，因此英文音素\f\的口型可以用汉语韵母\f\所对应的视位来表示。

在不考虑协同发音影响的情况下，每一个音素都有一个对应的视位表示。而实际上，有些音素的视位极为相似的，这些音素的口型可以用同一个视位表示，即音素和视位是多对一的关系，多个音素可能具有相同的视位。例如，\b\、\p\和\m\三个音素在发音时，其口型均为双唇紧闭，在视觉上极为相似，因而可以认为它们具有相同的视位。

英语音素无法直接映射到汉语视位，但是根据以上讨论，本文通过英语与汉语音素间根据发音部位、发音方式的相似性建立映射关系，进而利用已有的汉语音素的视位参数，最终得到英语音素的汉语视位表示。在建立英语与汉语音素映射关系时，首先根据文献[5]中的结论，得到英语音素根据发音部位和发音方式划分的分类，然后根据汉语虚拟说话人[1]对汉语音素的分类，二者进一步比较，最终将英语音素分为 15 类，对应到汉语虚拟说话人的视位参数表示。

2.3 基于分类处理的协同发音模型

根据文献[6]对英文发音口型的研究结论，在英文音素的发音口型中，元音音素的发音口型变化都是明显可见的，往往会影响与其相邻的音素的口型；部分辅音音素的口型是比较明显的，称为可视辅音，如\b\，\p\，\m\等；而其他辅音音素的口型变化幅度较小，称为不可视辅音，如\d\，\t\，\n\等，它们在连续语流中的口型主要受相邻的音素影响。

根据上述结论，各个音素的口型对相邻音素的影响以及受相邻音素影响的程度是不同的。本文根据音素对相邻音素的影响程度以及受相邻因素的影响程度，提出了一种基于分类处理的协同发音模型，它与传统的基于 Cohen 和 Massaro 提出的控制函数模型[3]相比，与具体的说话人无关，适用于可更换人脸模型的情况。

在本系统的实现中，所有英语音素被分为三类，分别是：1）可变可影响音素；2）可变不可影响音素；3）不可变可影响音素。这里的“可变”指的是当前音素会受到相邻音素的影响，导致视位发生变化；“可影响”指的则是当前音素会对相邻音素的视位产生影响。

根据上述分类，在本系统的实现中，协同发音情况被分解为以下两种情况：

1） 对于可变可影响音素：当其相邻音素为可影响音素时，根据该音素本身以及相邻音素的影响程度，将这些音素对应的视位融合为一个目标视位，以表示该音素受协同发音影响所形成的最终视位。视位变化可用公式（1）表示：

$$\tilde{V}_t = \alpha V_t + \beta V_m; \alpha + \beta = 1 \qquad (1)$$

其中 $\tilde{V}_t$ 是融合得到的最终视位，V_t 为当前视位，V_m 为相邻视位，α 和 β 为融合参数，根据视位的可影响程度决定：如果相邻视位对当前视位影响程度较大，则 β 较大 α 较

小；反之，则 β 较小 α 较大。

2） 对于可变不可影响音素：当其相邻音素为可影响音素时，该音素对应的视位使用其相邻的可影响音素影响程度最大的音素对应的视位表示。

实验证明，本文提出的基于分类处理的协同发音模型，针对连续英语语流中的协同发音问题，在一定程度上改善了因协同发音而导致的口型不自然、视位冗余等情况，具有较好的效果。

3. 基于 B/S 模式的 3D 双语虚拟说话人

所谓 B/S 模式，即浏览器/服务端（Browser/Server）模式。在这种模式下，用户工作界面在 web 浏览器中实现，用户通过浏览器访问指定网站即可访问网站上部署的应用。

本系统在实现中，采用了 ActiveX 技术将系统封装为浏览器插件，通过 IE 浏览器访问界面如图 2 所示。ActiveX 是微软提供的基于 Windows 平台的编程语言无关的可重用程序组件开发框架。使用 ActiveX 技术开发的控件，可以嵌入在网页中，通过 IE 浏览器访问和使用。使用 ActiveX 技术开发的控件可以访问几乎所有本地程序可以访问的操作系统资源，因此，利用 ActiveX 技术可以很方便的 Web 页面中添加与本地程序具有同样高质量、高交互性的内容和服务。

图 2 嵌入在浏览器中的虚拟说话人

通过利用 ActiveX 技术，本系统可以通过浏览器访问、使用，相比于传统的基于桌面的软件，具有更加易于获取和更新；可以调用 OpenGL 渲染库，实现 3D 渲染的硬件加速，与传统基于 Flash 的方式相比，具有更高的性能。

4. 系统评测

为了测试英文版虚拟说话人系统实现方法的正确性和效果，本文设计了一组主观评测实验，对基于视位映射的方法实现的英文虚拟说话人系统口型的音视频同步程度、自然度和准确度进行感知测评。

4.1 实验设计

测评实验随机选取 20 段长短不一的英文段落，作为评测的原始文本。

实验邀请了 5 位大学生作为被试者参加主观评测实验。实验中，被试者会先看到并理解每一段每一个要测试的样本数据的英文原文，进而，每一个样本数据会通过英文虚拟说话人生成相应的 FAP 并进行播放。被试者要通过自己的主观感受，对虚拟说话人系统生成的人脸动画口型的音视频同步程度、自然度以及准确度进行打分。打分采用 5 分 MOS（Mean

Opinion Score）评测方法，每一个分数对应的具体描述如下：1 分（效果很差，完全难以接受）；2 分（效果较差，对视觉造成干扰）；3 分（效果还可以，有部分错误，但可以接受）；4 分（效果较好，错误较少）；5 分（效果非常好，基本没有错误）。

为了主观评测结果更加准确，评测实验还设计了两个对比试验，一个对比对象是本文实现的英文虚拟说话人系统，但是没有使用协同发音规则；另一个是专门提供虚拟说话人服务的商业公司 Oddcast 开发的多功能的虚拟说话人系统 SitePal 系统[7]。在这两个对比试验中，为了降低其他因素的影响，测试数据均使用上述 20 个样本数据，但 SitePal 系统只需要提供音频文件，不需要标注文本文件。

4.2 实验结果与分析

根据以上的实验设计和实验数据，得到的实验结果如图 3 所示。

通过以上测试结果可以看出，本文实现的英文虚拟说话人系统在音视频同步程度和准确度上都与商用系统 SitePal 的效果相差无几，只是在自然程度上要相对差一些。这说明采用视位映射实现英文虚拟说话人系统的方式是可行的，而且具有较好的效果。另一方面，使用了协同发音规则的系统相比未使用的系统，在自然程度上分数有较大提升，说明本文采用的基于分类处理的协同发音解决办法具有较好的效果。

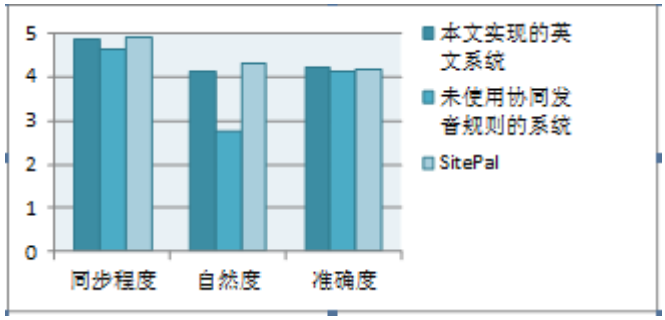

图 3 实验结果

5. 结论

本文采用视位映射的方法，通过将英文映射到汉语视位，建立针对英文发音特点的基于分类处理的协同发音模型，并利用 ActiveX 技术，将系统封装为可以通过浏览器访问、使用的 3D 双语虚拟说话人系统。通过主观评测实验和对比实验，证明采用视位映射的方法实现的虚拟说话人系统具有较好的自然度和准确度。

采用视位映射的方法，将汉语虚拟说话人扩展为一个英汉双语虚拟说话人是一个有益的尝试，将来可以考虑使用这一方法，将说话人扩展为统一参数的多语种虚拟说话人系统。

参考文献

[1] 王志明. 汉语视位建模及可视语音的研究. 清华大学，博士论文， 2003.4

[2] Ezzat T, Poggio T. MikeTalk: A Talking Facial Display Based on Morphing Visemes. Proceedings of the Computer Animation Conference, Philadelphia, USA, 96~102, 1998

[3] Cohen M M, Massaro D W. Modeling coarticulation in synthetic visual speech. Models techniques in computer animation, Tokyo Springer-Verlag, 139~156, 1993

[4] CMU, The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[5] 赵德梅. 英汉比较语音学, 青岛海洋大学出版社, 1995

[6] Jeffers J, Barley M. Speechreading, Thomas, Spring-field, IL, 1971

[7] Oddcast, SitePal, http://www.sitepal.com/

Research And Implementation Of Bilingual 3D Talking Head Based On B/S Structure

Lin Huijie+, Jia Jia, Wang Xiaohui, Lianhong Cai

State Key Laboratory on Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology（TNList）

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

+ Author: Tel: +86-13810358678, E-mail: linhuijie@gmail.com

Key words：viseme mapping; coarticulation; virtual talking avatar; B/S structure; ActiveX

Abstract：With the continuous development of computer technology, the intelligence and humanity of the computer is in great need. People hope software to be more user-friendly, intelligent and humanized. That means software must be easy to access, install and use. The technology of text to visual speech translation is an important part of human-computer interaction to make the computer more intelligent and humanized.

In this paper, the author describes an English-Chinese bilingual text-to-visual-speech synthesis system based on the Chinese 3D Talking Avatar platform. The English phonemes are grouped by viseme similarity and mapped to Chinese visemes to generate English version talking avatar. A rule-based coarticulation model is also implemented to solve the problem of coarticulation in English. Designed in B/S structure, the final synthesis system, with English and Chinese support, is released as a browser plugin, so users can easily access it via an internet browser.