www.ccw.com.cn
周报全文
首页 | 时讯与评论 | 知识中心 | 管理与职场 | e海航标 | I社区 | 计
2003年2月14日 综合新闻 / 产品与技术 / 应用与方案 / 网络与通信 / 市场与

文本-可视语音转换及其应用

王志明 蔡莲红

语音信号、视觉信号和文字是人类信息和知识的主要载体，也是人类进行学习和交流的重要工具。在电子和通信技术迅速发展的今天，多种媒体之间的交互作用越来越受到人们的重视，如语音合成(text-to-speech)与自动语音识别实现了文字和语音的互相转换；自动机器唇读、图像辅助语音识别和音视频联合编码则利用了语音和图像之间的内在联系。

经过数十年的发展，语音合成技术已经走入实用阶段，在信息咨询、电话银行、车站播报系统等各个方面有了广泛的应用。近年来出现了另一种多媒体研究热潮，即把声音和文字、图像集成在一起，形成直接由文本到可视语音的转换(text-to-visual speech，TTVS)，使人们在听计算机说话的同时能看到一个合成的人脸，使人机交互界面更为友好、和谐。

对视觉语音（visual speech）的研究正是这样一种综合考虑声音和图像的多媒体技术。视觉语音是指人们在用语言交流时所表达出的面部表情和动作，它能在一定程度上传达人们想要表达的意思，并能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”（talking head），即表现说话者面部表情和嘴部、眼部等变化情况，则会大大改善人们对声音的理解。在人机交互的过程中，如果人们面对的不是单纯的文本，而是一个会说话的人物形象，则使人觉得计算机界面更为友善，方便人们与计算机的交流。近几年来，对视觉语音的研究越来越受到人们的重视，已成为多媒体和人机交互技术研究领域相当活跃的研究方向。

TTVS的实现

对于TTVS，其实现方法可分为以下两类：

基于参数控制的方法 首先对人脸建立一个网格模型，包括多个多边形(一般是三角形)和顶点。由一组参数来控制每个顶点的运动，再通过图像变形技术实现人脸上各个像素点的运动，来生成人们说话时的各种面部表情。该方法的优点是需要的数据量小、控制灵活、可移植性强;缺点是合成的图像往往带有人工制作的痕迹，但对于这一点，各国研究者正在努力改善。

基于数据驱动的方法 类似于语音合成中的波形拼接合成法。通过对人们说话时可能出现的各种表情进行录像，从中提取大量的原始数据，建立图像数据库。在合成时从库中选择合适的图像进行拼接，并进行一些消除图像边缘效应和抖动的处理，生成动态的连续的说话者的面部表情。该方法的优点是合成的人脸图像质量高，较为逼真、自然;缺点是在建立模型的训练阶段需要大量的原始数据，生成的数据库需要保存大量的图像数据，且所有数据完全是针对某个特定人的，无法移植到其他人身上。

现在运行的系统中多为参数控制系统，其中控制参数也多采用MPEG-4所定义的人脸动画参数(facial animation parameter,FAP)。MPEG-4制定了一整套人脸模型化描述方法，包括用于定义人

脸模型的面部定义参数(facial define parameters,FDP)和一组用于定义人脸面部动作的人脸动画参数FAP。其中FDP通过对人脸上84个特征点的位置信息来定义人脸模型，这些点不仅包括外表看得见的人脸特征点，还包括了舌头、牙齿等口腔内器官的特征点，如图1所示。

FAP一共有68个参数，包括两个高级参数和66个低级参数。高级参数是视位(viseme)和表情(expression)，视位分为15个，分别表示人们发某一音位时的面部动作;表情分为高兴、悲伤、愤怒、害怕、厌恶、惊奇六种。66个低级参数用来控制部分FDP特征点的运动，进而形成各种复杂的人脸动作。这些标准的制定极大地推动了参数控制合成方法的发展，使这种方法在人机交互、计算机网络交谈、游戏动画等方面得到更为广泛的应用，图2 是参数控制的TTVS系统的基本框架。

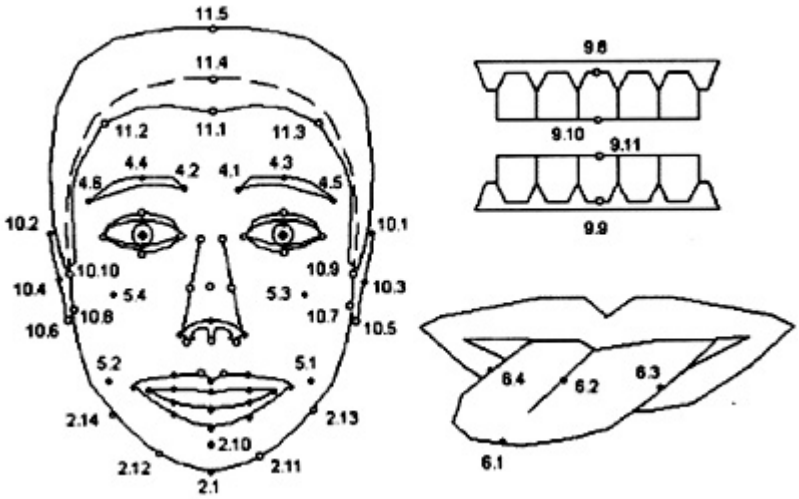

图1 MPEG所定义的FDP特征点

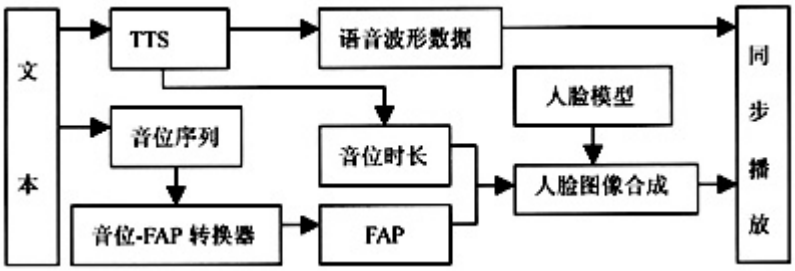

图2 TTVS系统的基本框架

VSonic系统

目前，国内外众多研究机构和公司均十分关注TTVS的研究，如MIT、AT&T、Microsoft、Motorola等。目前，清华大学计算机系已经开发出了具有自主版权的汉语TTVS系统。

清华大学计算机系致力于人机语音交互的研究始于1979年。在20年的研究中，得到了国家自然科学基金、国家863计划、国家重点攻关项目和军事预研项目的资助，取得了一系列国内外领先的研究成果，并多次获奖。在语音合成方面，我们深入地研究了声学模型、韵律模型、文本分析、韵律描述语言等语音合成中的关键技术，并于1993年推出了汉语TTS软件产品。1999年实现了基于数据驱动的汉语TTS系统Sonic，获得了高自然度的语音输出。

为了增强TTS系统界面的友善性，清华大学计算机系于2000年着手研究汉语语音的可视化，为其原有的Sonic系统配上发音人的头像，形成了新的汉语文本-可视语音转换系统VSonic，系统界面如图3所示。

图3 VSonic系统运行界面

在VSonic系统中，人脸模型是一个由三角形组成的二维网格人脸模型，整个模型共包括约220个点和350个三角形，如图4所示。模型中的顶点涵盖了由MPEG-4定义的主要FDP特征点，模型的驱动参数是标准的FAP参数。人脸合成是以单一的真实人脸正面照片为基础，在FAP参数的控制下对人脸图像进行变形处理(warping)，首先求得FDP特征点的运动向量，再通过其余点与这些点的位置及拓扑关系求得模型中所有顶点的运动向量。根据顶点的运动向量和对三角形的平面近似，利用双线性插值方法求得所有像素点的运动向量，从而使人脸“动”起来。对于口腔内的图像，我们采用固定的模型，具有真实的牙齿和口腔内图像纹理，并能根据开口度的大小和上下唇的突出度来调整亮度。

系统由语音合成部分提供时间同步信息，实现完全同步的语音和图像播放。系统中语音的发音速度可调，图像以固定的帧速率播放，不受语音快慢的影响。当语音速度加快时，每个音节的图像帧数将减少；反之，当语音速度放慢时每个音节的图像帧数将增加。图像的帧速率可根据系统性能来调节，使系统在各种性能的机器上均能保持语音与图像的同步。

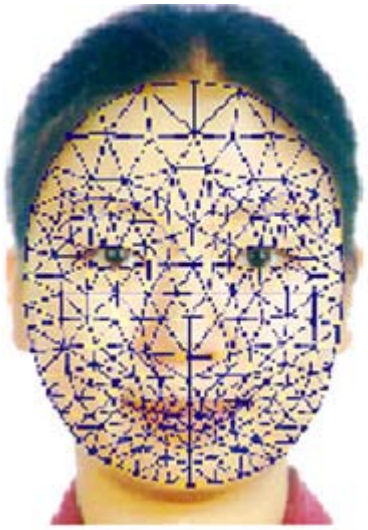

图4 人脸网格模式

除了能够生成各种各样的说话口形外，VSonic还能表现出眨眼等简单的面部动作，以增强系统的自然性。系统的另一特点是其可移植性非常好，可以使人脸模型从一张人脸更换到另一张人脸。

只要有一张正面人脸照片，借助系统提供的工具，经过简单的鼠标操作即可在数分钟内实现系统中人脸模型的更换。

TTVS的应用和展望

文本-可视语音转换系统不仅提高了人机交互界面的友善性，丰富了人们的生活，还在许多领域中有着重要的实际意义。下面我们介绍几种TTVS在实际生活中的应用。

（1） 制作虚拟电视节目主持人

这是TTVS一个很好的应用实例。虚拟电视节目主持人在许多国家已经走上了屏幕，引起了广大观众的极大兴趣。图5是英国报业联合通讯社推出的第一个虚拟新闻播报员“ 阿娜诺娃 ”(Ananova)。

（2）增强语音的可懂性

实验表明，在噪声环境下，能看到说话者的人脸相当于提高了8～12dB的语音信噪比。因此，在环境噪声较大的情况下，如在工厂车间、高速运行的交通工具上或战争前线进行人机交互时，如果在机器给出语音的同时能给出一个合成的人脸，则能大大改善人们对语音的理解。另外，在听话者有听力障碍的情况下，也有类似的效果。

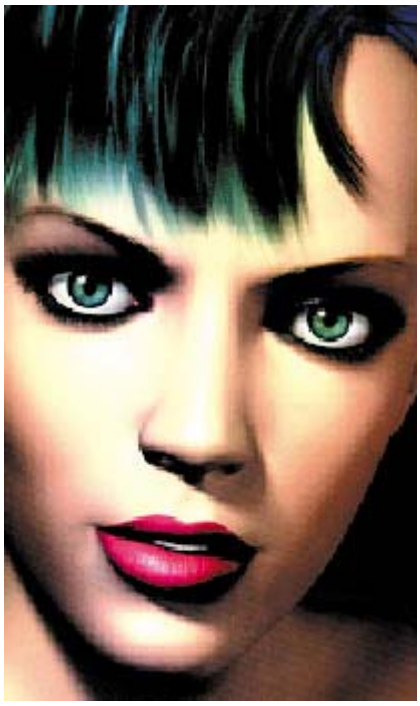

图5 Ananova

（3）网上聊天

现在网上聊天主要是通过窗口中的文本进行交流，如果人们在网上聊天也可以像实际生活中聊天一样，既可以听到声音，又可以看到说话者的人脸，将会大大增强使用者的兴趣并方便交流。但现在网络带宽不能满足实时传输声音和图像数据的需求，如果在用户的计算机上安装了TTVS系统，则可以在网上只传送文本信息，而在本地由TTVS合成语音和图像，使用户既听到声音又看到说话者的人脸。若在文本中再加入少量的标注信息，还可以使人脸表现出各种各样的表情。再进一步，如果在用户的计算机上安装上话筒和相应的语音识别软件，则用户可以脱离键盘，就像日常生活中一样，与对方面对面地聊天。

另外，在越来越广泛的商业、娱乐人机交互的过程中，如新产品介绍、电子游戏等，如果人们面对的不是单纯的文本或声音，而是一个会说话的人物形象，则使人觉得更为亲切，更容易接受，从而提高商业销售额，给企业带来巨大的经济利益。

总之，TTVS技术的出现是多媒体技术迅速发展的产物，也迎合了社会发展的需求。它给人们的生活增添了新的色彩，使计算机更人性化，人们与计算机的交流变得更为简单。相信在不久的将来，它将会在众多的技术、商业和娱乐领域得到广泛的应用，并逐步进入我们每个人的生活。