面向自然交互的多通道人机对话系统

杨明浩，陶建华，李昊，巢林林

中国科学院自动化研究所模式识别国家重点实验室，北京 100190

摘　要：人们在对话过程中，除了使用口语交互外，还会很自然地利用表情、姿态等多模态信息辅助交流。本文重点分析并阐述了如何将这些多模态交互方式有效的融合到人机对话模型中，并实现一个面向自然交互的多模态人机对话系统。首先根据不同通道（如情感、头姿）对语音交互的影响，将它们主要分为信息互补、信息融合和信息独立三种模式，并针对三种模式分别采用不同的方式实现输入信息的多模态融合。信息融合后的对话管理，采用有限自动机、填槽法和混合主导方式的对话管理策略。针对对话中的情感处理，提出一种情感状态预测网络记录用户的情感变化，并根据话语的轮转的不同对话上下文对用户情绪变化进行及时反馈，该对话模型能比较灵活地处理用户在对话过程中呈现的多模态信息。信息输出方面，针对人机对话中较为常用的数字虚拟人的行为控制，提出了一种简化的多模态协同置标语言，实现了虚拟人的包括情感、姿态与语音的同步表达，提高了虚拟人的表现力。基于以上关键技术，最后实现了一个面向城市路况信息查询的多模态自然人机对话系统，相对于传统的语音人机对话模型，多个用户的体验表明本文的多通道自然人机对话系统能有效提高用户交互的自然度。

关键词：多模态信息融合；人机交互；对话管理

1．引言

自计算机问世以来，人类就梦想着有朝一日能与计算机进行自然的对话，便捷的获取计算机提供的各种服务。近年来，随着语音识别、语音合成以及数字虚拟人表达技术的发展，人与计算机的自然对话已经获得很大的进步，如英国 BBC 电视台的网络女虚拟主播 Ananova[1]，日本名古屋工业大学的数字虚拟人等等[2]，美国南加州大学的数字智能生命体（Creative Agent）[3]，这些虚拟人能以逼真的语气朗读用户给定的文字，理解用户的查询需求，回答用户的购物问题和票务信息查询系统信息等等，甚至还可以以幽默的口气对语音识别不准确的问题进行反问，如苹果公司的语音助手 Siri。可以说数字虚拟人与人的自然对话已经在实验室环境下取得长足的进步，成为自然人机交互的重要发展方向。然而，目前的自然语音交互技术距离实用化以及进入人们的生活，还有很多问题需要解决，其中一个重要的方面就是人与计算机的对话很不自然，如计算机对交互过程中人的情绪、姿态和语气变化缺乏良好的反馈，目前的订票、旅游信息查询等人机对话系统中，当用户对查询结果不理解或者没有得到满意答案时，系统通常缺乏对用户状态的积极反馈，使得用户的体验较差，不愿多次使用。另外，人机对话的输出比较呆板，如目前的大多人机系统多采用语音合成或者数字虚拟人的方式输出对用户问题的回答，通常数字虚拟人的动作都事

资助项目：中国自然科学基金（项目批准号：61273288, 61233009, 61203258, 60873160, 61011140075, 90820303）

联系作者：杨明浩，E-mail：mhyang@nlpr.ia.ac.cn

先录制好，当对话的回合较多，数字虚拟人的动作就会重复，也在一定程度上降低了人机对话的自然度。

为了解决这个问题，本文提出了一种面向自然交互的多通道人机对话系统原型，这里自然交互指对计算机对用户在对话过程中的情绪、语气、姿态（如头姿）等变化进行实时的检测和跟踪，同时，为了确保数字虚拟人能较好的对用户交互进行反馈，我们根据不同通道对语音交互的影响，构建了信息互补、信息融合和信息独立三种信息融合模式，来实现输入信息的多模态融合。信息融合后的对话管理，采用有限自动机、填槽法和混合主导方式的对话管理模型；针对对话中的情感处理，提出一种情感状态预测网络记录用户的情感变化，并根据话语的轮转的不同对话上下文对用户情绪变化进行及时反馈。信息输出方面，针对人机对话中较为常用的数字虚拟人的行为控制，提出了一种简化的多模态协同置标语言，实现了虚拟人的包括情感、姿态与语音的同步表达，提高了虚拟人的表现力。

本文后面组织如下，第二节首先介绍相关工作；第三节给出本文提出的多模态人机对话系统框架；对话管理模型中多模态信息融合方法，对话中的用户情感信息处理策略在第第四节介绍；第五节介绍虚拟人的多通道动作协调控制方法；最后，面向城市路况信息查询的多模态自然人机对话系统以及总结分别在第六节和最后一节讨论。

2．相关工作

早期，研究者们提出了很多面向自然人机对话的原型系统，这一阶段的人机交互多侧重于单一模态（或称为单一通道）的信息处理，如：以语音识别和合成为基础的口语对话系统、人脸表情跟踪与识别系统、手势识别与交互等。然而在人们的对话过程中，当一个人的语音或语气不足以反应具体表达的意思时，有时能从脸部表情或肢体动作上判断出说话者意图，甚至一个简单的表情，辅助伴随的手势动作快与慢、幅度变化也会蕴涵丰富的交互信息，可见多模态的人机交互方式在表达效率和完整性上都要优于传统的单一模式。因此后续的研究都侧重在多模态自然人机对话上，其中最重要的一个研究内容就是对话管理的研究。

对话管理技术在早阶段的一个重要目的是降低计算机对语音识别文本的理解错误，让计算机正确理解用户的提问，完成用户指定的操作[4]。这个阶段的对话管理系统，用户问题集通常相对简单，对话逻辑可预先预测，多采用基于规则的方法进行构建，如填槽法[5,6]，有限自动机方法[7]等等，这类以规则为主导的人机对话模型在商业上获得了成功的应用，如到现在为止，基于规则的对话管理技术依然在呼叫中心等计算机电话自动处理业务中大规模使用。近年来，随着语音识别技术和语音合成技术的发展，越来越多的机构和研究单位更加重视自然的人机对话研究，如美国 DARPA 计划、欧盟框架计划、日本 JSPS 计划以及我国 863 计划和自然科学基金长期以来均在此领域设立了相关研究项目。这一时期，随着计算机可以快速处理大规模数据，基于统计模型的对话管理技术，如贝叶斯网络[8]，图模型[9]，基于对话的增强学习技术[10]，部分可观测的马尔科夫决策过程(POMDP) [11]等等，使得计算机能够灵活的处理人机对话过程中用户的输入错误，相对于传统的基于规则的对话模型，基于统计模型的对话管理技术给与了用户在对话过程中较大的自由度。然而也由于这样的自由度，使得统计方法的计算复杂度较高，如 POMDP 模型中，对话状态中

信念状态数增长，会给计算复杂度带来指数级增长[12]，尽管一些加速技术的提出，在一定程度上降低了时间复杂度，但由于多模态对话管理过程要综合考虑来自语音、表情、姿态等多种信号的融合，因此完全基于统计模型人机对话系统依然较难用于实际的人机交互中。因此一些研究者结合规则与统计模型两种手段来建立人机对话管理系统，在保证对话精确度的同时，尽量减少对话过程中的计算复杂度，以期望能构建可以面向实用的人机对话系统[13-15]。

在对话的话题轮转方面，谁主导对话的过程是确保对话过程流畅的关键，谁主导对话过程也与多通道信息融合的方式密切相关。根据主导人机对话进行的角色不同，主要可分为系统主导模式，用户主导模式和混合主导模式三种对话控制策略[16]。不同模式处理对话过程有一定不同，如在系统主导模式下，系统需要实时针对用户提问或者行为变化作出反应；而在用户主导模式下，系统则稍微隐蔽的监测和跟踪用户的行为，综合分析一段时间内的用户行为，在恰当的对话轮转时给出应答。这方面，典型的方法有长短时记忆递归神经网络（“Long Short-Term Memory Recurrent Neural Networks”），该方法分别针对短句和长句的情感累积变化，来对对话过程中用户的情感变化[17]做出反馈。近年来，随着语音识别、语言理解和对话错误侦测技术的提高，虚拟人已经具有快速反馈用户语音命令的能力[18]。然而无论在哪种模式下，系统除了识别语音，还需要检测和跟踪用户的情感和姿态变化。所以如何针对更多的体现用户意图的用户行为，做出及时的人性化反馈，是人机自然对话的一个重要研究方向。

数字虚拟人对用户行为给与及时反馈，能带给用户良好的交互体验。早期的数字虚拟人表达技术多集中在动作产生，实时动画生成，以及如何让虚拟人更有表现上的研究上。后来，数字虚拟人表达技术的研究主要集中使虚拟人的多模态动画与语音的协同控制上，如行为规划法[19]和行为标记语言（Behavior Markup Language (BML)）[20-22]等等。一个良好的数字虚拟人控制接口不仅要使不同模态的动画与语音进行良好的同步外，还要格式简洁灵活，易于编写，这样才能很好的与对话管理模块结合，使数字虚拟人足够“聪明”，并很好的与用户交互。

本研究针对目前的多通道自然人机语音交互现状，给出了一个面向实用的多模态自然人机语音交互对话模型。相对于传统的人机对话模型，本研究工作的创新点主要在于：（1）有效的将用户的多模态交互行为方式（包括用户的语音信息、情感信息和姿态信息）融合到多模态人机对话模型中；（2）针对人机对话中较为常用的数字虚拟人的行为控制，提出了一种简化的多模态协同置标语言，实现了虚拟人的多通道情感动作表达和语音协调控制，提高了虚拟人的表现力；下面详细描述。

3. 系统结构

与其他多模态人机口语对话方式一样，本研究提出的多模态人机对话包括信息获取、信息处理、信息输出三个部分，如图 1 所示。信息获取模块通过麦克风、摄像机等输入设备接收来自语音、表情、姿态等通道信息，然后借助多模态信息分析融合模块，产生多模态协同对话内容，并同步输出到系统显示设备。

本研究与其他多模态人机口语对话系统不一样的地方主要在于：更注重对人机对话过

程中用户行为变化的反馈。我们的多模态信息分析与融合模块，强调对用户行为的融合与分析，包括针对用户情感、头姿变化的跟踪与判断，并与用户的语音信息有效的结合。在多模态信息表达表达部分，为了提高用户的自然体验度，建立了一个简单但灵活的虚拟人控制接口，与语音驱动的唇形、表情动画一起，实现了具有多种情感表达方式的数字虚拟人合成。图 1 中灰色的矩形区域是本系统中的主要模块，其中多模态信息融合和对话管理模型，用于处理对话过程中不同通道信号在不同时序上的融合，及以语音为基础的多模态对话控制策略，是提高多模态人机对话的自然度的关键。多模态虚拟人表达模块主要使对话管理模块与多模态数字虚拟人表达产生良好的衔接，通过协同置标语言能将虚拟人的动画与语音输出信号协同表达，使数字虚拟人很好地对用户进行反馈。

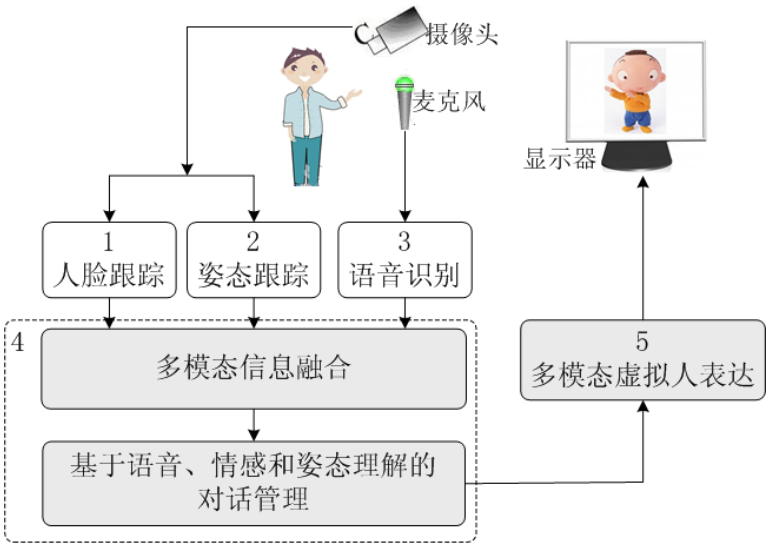

图 1　多模态人机对话管理框架

3. 多模态融合的对话管理框架

3.1 多模态融合策略

多模态融合，是对话管理模块根据用户语音、表情、姿态的识别结果，对其意图进行理解的前提。由于人机对话中通常以语音信号为主，因此，本文根据不同通道对语音交互的影响，根据它们与语音信号的关系把它们的处理方式分为三种模式：信息互补模式、信息融合模式和信息独立模式，三种模式的处理方式如下：

在信息互补模式下，对话过程中的语音内容必须与其它通道信息作为补充才能构成完整的语义，如：当用户说“我想去那边”，对话系统需根据手势所指方向，或者根据人脸所面对的方向，来判断到底用户询问的具体方向，然后才能做出反馈。

在信息融合模式下，每一个通道均表示相似的语义信息，甚至有时是可以相互替代的。如：“点头”和语音中的“是”、“可以”等内容是表达的相同语义；同样，竖起大拇指和语

音中的“很棒”也是可以相互替代的。这种情况下，每个通道都能表示完整的语义，它们可以单独工作，也可以同时出现以获得比单一模态语义分析的更好的表达效果。信息融合模式还有一种特殊情况，就是多个通道信息在时序特征上结合特别紧密，它们之间存在着时序上强耦合的关系，如：语音和唇动、说话语气和脸部表情等等。因为在这种情况下，其它通道的信息与语音信息在时序上具有较强的相关性，可以采取相对快捷的特征层融合思路进行多模态融合处理，这种处理方式多用于多模态情感识别或者多模态语音识别。

在信息独立模式下，每一个通道表示的语义信息相对独立，它们之间没有很强的约束关系，但有时可以增强其它通道的表达效果，如：当人说话时一些手势或头姿运动，有时可以起到增强语气表达的作用。在这种模式下，语音通道在语义理解中起主要作用，其它通道的信息多用于情感判断，以达到更好的理解用户意图的目的。

这几种不同的多通道融合方式考虑了不同通道信号在语义上的关联性，实现了在不同层次上的融合处理，能有效提高人机对话模型的自然度。由于多模态对话过程中，并不总是需要语音，其它通道也会产生具有语义表达效果的信息，因此，在实际的对话系统中，本系统的融合方式采用如下方式进行：

（1） 对话过程中，系统同时接收并记录用户语音、头姿和手势等多个通道的信息输入；

（2） 对话过程的某一时刻，只有语音信息输入时，系统直接根据语音内容做出对话响应；

（3）当不同通道同时有信息输入，系统首先分析语音的内容，如果语音内容对语义的理解不构成歧义，则直接根据语音内容做出对话响应。如果语音内容还需要其它通道信息作为补充才能构成完整的语义，则结合其它通道信息，根据信息互补模式下的工作方式进行综合判断。如：当用户说“请将这边的内容告诉我”时，对话系统会根据手势所指方向，来判断需解释的内容，并做出信息反馈。

（4）如果在一段时间内没有语音信息输入，系统根据近一段时间的用户脸部表情、头姿或者手势变化历史记录，根据这几个通道的信息判断用户的语义表达。如：用户“点头”或者举出“OK”手势时，表示同意；而“摇头”或者“摆手”则表示不同意；在不同的上下文环境时，“摆手”又可以表示再见的意思。而当用户在对话过程中，停止说话，同时肢体行为缓慢，脸部表情处于类似于“悲伤”或者“安静”状态时，则有可能实际处于“思索”的状态，在这种情况下，系统根据当前对话过程的上下文内容进行一些提示性的询问，要求用户进行反馈。

3.2 多模态对话管理

在多模态融合语义理解的基础上，需构建对话管理模型以确定系统的应答方式。对话管理模型决定着对话状态的转移和上下文语境下的应答，它在很大程度上决定了系统的表现和行为，如何设计好对话管理模块是保证人机对话过程自然连续的关键。对话管理模型的设计主要包含两个方面：对话主导策略和对话控制策略。

由于本系统的对话过程需要考虑来自语音、头姿和手势等多个通道的信息融合结果，除了系统应该能回答用户提出的问题外，而且在某些情况下，需要系统能主动提出问题来澄清一些模糊的概念，这些情况包括语义理解错误、用户情感变化、姿态含义的模糊、提供信息不足、上下文语义不一致等等。这种情况下，混合主导方式能较好满足这种对话需

求，即本系统中，用户和系统都能掌握对话的控制权，提问或者发问，对话过程中的对话控制权是随着对话过程改变的。混合主导的人机对话模式比单一系统主导或者用户主导的对话方式更有效更灵活。

对话管理的另一个重要步骤对话控制策略，其主要功能是通过一定的控制策略，推进人机对话自然合理的进行。本系统采用有限自动机结合填槽法来实现混合主导的人机对话管理，并将多模态融合的三种模式有机的融入对话过程中。整个对话中的状态对应于有限自动机的三个节点，分别是：询问、回答和漫谈，对话过程就在这三种状态之间跳转。当对话系统应用于问题查询时，用户通过不同的查询命令，以及不同的情感状态，促使系统在三个状态之间跳转，这部分为用户主导。在每一个状态内部，系统设置了一系列槽，并采用填槽法进行管理，当用户的输入信息不足以填满槽时，对话系统就产生疑问，或给出反问语句要求用户进行回答，并等待用户的语音、表情、头姿或手势等通道信息输入。这里，系统分别根据信息互补模式、信息融合模式和信息独立模式，依照多模态融合策略获得用户的语义信息，然后对槽内容进行填充。如果槽填满，则进行状态转移，并根据槽内的信息，对用户提出的问题进行解答，这部分为系统主导。解答信息后，系统会根据用户不同的姿态和情感输入进入下一轮讯问状态或者根据用户情感的变化转移到漫谈状态。图 2 给出了本文人机对话管理模型，这种以有限自动机和填槽法为基础，采用混合主导方式的多模态对话管理机制，能比较灵活地处理用户在对话过程中呈现的多模态信息，实现比较自然的人机对话过程的控制和管理。

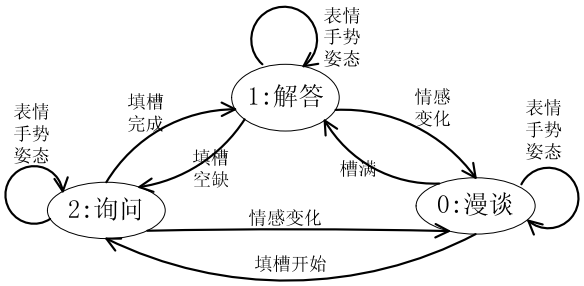

图 2 对话管理自动机跳转图示

3.3 对话管理中的情感处理

情感作为人机对话的重要组成部分，对交互过程起推动和辅助作用。一个愉快的对话过程同时也是一个情感的交流过程，因此情感识别和理解在自然人机对话过程中起非常重要的作用。在本文中，情感在对话系统中的作用方式分为三种形式：

（1） 对语义理解起到互补作用，情感是语义理解的一部分，如：疑惑的语气表示疑问状态等。在这种情况下，情感将被直接应用于对话系统的语义理解；

（2） 用户产生负面情绪时，将导致对话系统转移到系统主动询问状态，并采取一些特殊的用语，以安抚用户的情绪；

（3） 当用户产生明显的正面情绪（如高兴）时，系统会短暂的进入漫谈状态，以活跃现场气氛。

在自然人机对话过程中，用户的情感识别是一个较为复杂的过程，可以从用户的语音

内容、语气、脸部表情和部分手势等参数进行综合判断。本文针对不同通道下的情感识别采用下列方式：

3.3.1 从对话语音内容上判断情感

针对语音内容的情感识别，是指在语音识别的基础上，通过分析语音内容获得用户希望表达的情感，由于在这种模式下，情感与语义密切相关，完整的分析语义具有很大的困难。为此，针对文本内容的情感预测中，主要放在语音识别后的文字所表达的正负情感极性预测上。

针对对话语音的特点，本文提出了一种情感状态预测网络(Emotional Status Prediction Network, ESiN)。它的核心思想是利用情感关键词来进行情感的判断。情感状态预测网络的初始步骤，首先是确定情感的焦点。情感焦点在通常情况下，由情感关键词驱动，多出现在情景对话和具有剧烈变化的情感状态中。在发怒语气中，承载发怒的情感关键词得到了突然加强，而“激动”的情绪能够通过加强功能词或情态词而得到明显的表现。例如：“我非常生气”中的短语“非常生气”表示了句子的关键的情感状态，并且在愤怒的情感时会得到有力地加强。其它的一些词语，如：“不好”，“很”，“非常”等等也会达到同样的效果。

ESiN 网络的基本组成部分是结点和链路。结点（情感载体）可承载情感信息而链路（情感传播者）可传播情感信息。ESiN 网络中的每个结点中有三个属性：词语（包括属性，如：词类、情感属性等）、情感矢量（$\vec{E}_t = \{e_{t0}, e_{t1}, \cdots, e_{tN}\}$），t 表示不同的节点。情感矢量表征一个节点的基本情感状态。情感矢量的分量分别代表不同的情感状态，N 表示情感状态的个数每一个分量的值从 0 至 1 分布，表示该情感状态的程度。其中 1 为最高，0 表示该情感状态不存在）、情感触发器（用以综合情感矢量，用以计算到当前节点为止的情感状态）。

ESiN 的链路表示情感的传播路径。根据不同语法规则，网络中有不同类型的链路。情感可以无延时传播，每个链路包含三个信息：方向（链路的起源和终结）、情感延时函数（延时函数控制情感传播）、情感生成概率函数（用于决定情感激励状态的概率）。

ESiN 网络如图 3 所示，整个工作过程包括：情感初始化、情感触发和情感传播等过程。情感初始化包括：分析语句中是否存在情感关键词并确定情感焦点等步骤，同时确定网络初始节点的情感矢量值 $\vec{E}_0$。

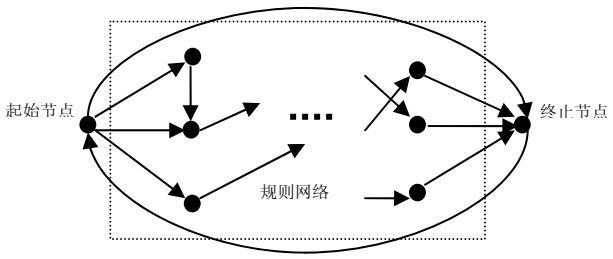

图 3　ESiN 情感状态预测网络

在情感传播中，若一个语句时段中包含情感关键词（情况（a）），这个时段的情感判

别结果就用作传播源。若某时段没有传播关键词没出现，我们就寻找功能词等信息（情况（b））。若情况（a）和情况(b)都没发生，则有情感触法值的词被用作传播源。这构成网络的初始化过程，情感传播值 Ep 从传播源计算并用转换延时函数(1)与情感矢量结合起来。延时函数(1)定义来确保情感适于传播。为了确保情感的汇集，情感传播在经一些阶段后延时为零。据以上标准，我们定义以下函数：

$$\vec{E}_t = D(\vec{E}_{t-1}) = \vec{\delta}(t) \times \alpha + \vec{E}_{t-1} \exp(-.005 \times t^2) + \vec{C}_t$$

其中：

$$\vec{\delta}(t) = \begin{cases} 1 & \text{，当前节点为情感关键词} \\ \{P_0, P_1, \cdots, P_N\} & \text{，当前节点为非情感关键词} \end{cases}$$

$$P_n = P(O_t \mid (O_{t-1}, O_{t-2}, \cdots, O_{t-M}), n)$$

E_{t-1} 表示在与当前节点相关的前一个节点情感矢量。$\delta(t)$ 表示当前的情感激励，若当前节点为情感关键词是，则输出一个新的情感激励源 1；若当前节点为非情感关键词时，则通过情感生成概率函数，确定当前的情感激励状态，其中 O_t 为节点 t 的词类标注。α 表示情感抑制系数，用以表示情感激烈程度。C_t 表示情感矢量的修正值，用以人为修正当前节点中的情感矢量值，主要是针对一些特殊的情感用语需要做一些调整。在正常情况下，情感矢量 E_t 在失去持续激励的条件下，经过一定的节点，将逐渐衰减为零。

通常情况下，在 ESiN 网络的终止节点中，情感矢量只有一个分量不为零。这表示，在一般情况下的情感状态通常由主要的情感关键词来决定。但在，有些场合，会出现情感状态歧异的情况，这主要由情感关键词、标注本身出现多意性导致。情感触发器的目的，是将多个不为零的情感矢量分量根据其节点的历史记录进行综合，得到确定的到当前节点为止的情感状态。

$$M_t = \arg\max_n (\sum_{i=0}^{t} e_{n,i})$$

从语音内容中分析得到的情感，往往与语义密切相关，对话系统将语音内容中获得的情感状态按以下两种情况处理：

（1） 如果情感的状态与用户想要提供的信息（即填充槽所需信息）密切相关，则被用来做语义理解。

（2） 大部分的情况，情感识别的结果只是反应了用户的一种情绪状态，如在天气查询的对话应用中，当用户说出“天气真糟糕”时，对话系统可以根据用户的这种负面情感，提供一些积极的信息，如第二天的天气会好转时，系统可以说“不用担心，明天天气就会好转”，以达到更为自然的人机交互目的。

3.3.2 从语音语气和表情上判断情感

除语音内容外，人机对话过程中的情感还往往同时隐藏在语音的韵律（语气）和人脸表情中。由于语音韵律和脸部表情的关联性较强，因此同时融合语音和表情的情感识别，通常采用信息相容模式下的多通道融合方法。在这里，本文采用了基于 Boosting 算法和分

类回归树(Classification And Regression Tree，CART)对来自语音和人脸特征点分布模型的双模态情感识别模型[23]（如图 4 所示），采用该方法能够动态的调整情感分类器的权重，实现对语音韵律参数和表情参数不同特征的权重进行较为合理的分配，从而能有效提高情感的识别准确率。

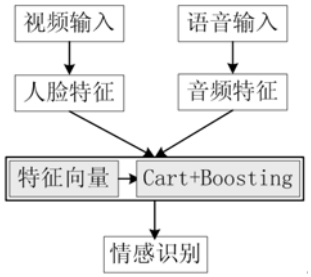

图 4 基于 Boosting 框架的多模态融合和情感识别流程

3.3.3 不同通道情感信息模糊或者矛盾的情况

有一种特殊的情况，就是语音通道所表述的情感，和其它通道表述的情感完全相反，如：当用户说出赞扬话语“你说的很好”时，脸部表情却流露出生气或者悲伤的情感，通常是表述的相反语义，甚至是挖苦的意思。在实际的多模态对话系统中，当不同通道所表达的情感不同时，可以采取极性相乘的原则，即首先将每个通道的情绪规整到正负极性上，然后将各个通道的情绪极性相乘，获得最终的情绪极性。在这种情感判断模式下，只要有一个通道的情感是负面的，则最终输出的情感判断就是负面的。

4. 多模态对话中的虚拟人表达

4.1 虚拟人行为的多通道协同控制

数字虚拟人作为多模态人机对话系统的输出模块，提供给用户自然的交互感受。针对来自不同通道的融合信息，虚拟人需要在说话时，做出不同的动作、手势以及表情，由于对话过程的灵活多变，很难采取事先动作定制的方式，来实现虚拟人的表达。因此，如何在虚拟人表达时，实现其行为的协同控制，并能与语音进行完美同步，是一个很重要的技术难点。针对此项工作，W3C 分别制定了 EMMA（Extensible Multi-Modal Annotations）[24]、SSML（Speech Synthesis Markup Language）、EmotionML（Emotional Markup Language）[25]等协议标准，它们采用 XML 框架，分别对多模态内容、语音合成和情感内容进行标注和控制。综合使用这些协议标准，可以较为有效控制虚拟人的动作、表情和语音输出的方式。本文作者实际参与了 EmotionML 和 SSML 两项标准的制定，本对话系统在此基础上，针对多模态虚拟人表达的特点，设计了一个简化的多模态协同控制语言(ML：Multi-modal Language)，分别在高层参数（如动作、姿态）和低层参数（如语音、唇型、头姿）上实现了虚拟人的多模态特征融合。ML 的结构如下所示：

<emotion=** animation=** degree=**>文本段 1<emotion=** animation=** degree=**>

文本段 2<……>

其中，"文本段 1"或者"文本段 2"是虚拟人需要表达的语句。关键字"animation"为虚拟人在说话时需要表演的动作，包括常用的抬手、打招呼、摆手、鞠躬、走路、跨步等动作片段。"emotion"为虚拟人动作时对应的表情，其为六种基本表情中的一个（中性、高兴、悲伤、愤怒、惊讶、害怕）。degree 代表虚拟动作以及表情的强烈程度，分为强弱两个等级，图 5 展示了高兴与愤怒两个动作的不同强弱程度的虚拟人动作情况。在多模态虚拟人数据库中，我们给出了 6 种基本情感状态下，强弱程度两种不同等级的 42 个身体动作，共 84 个动作序列，能表达出常见的类人姿态和表情。虚拟人表演动画时，不同动画片断之间的衔接采用插值平滑方式完成，保证了虚拟人动作的连续性。

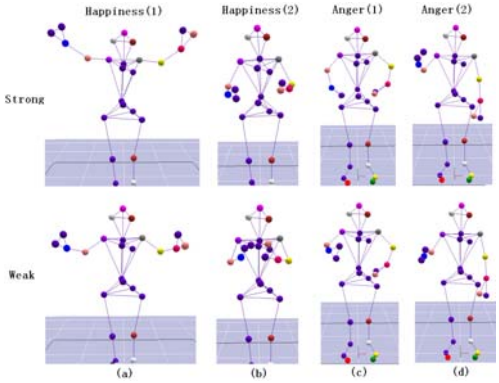

图 5 虚拟人同一情感的不同强弱程度的表达

图 6 完整的表示了虚拟人以协同控制语言作为输入，根据 ML 的内容解析结果，产生基于动画，实现唇形、脸部情感、头动和姿态控制的过程。

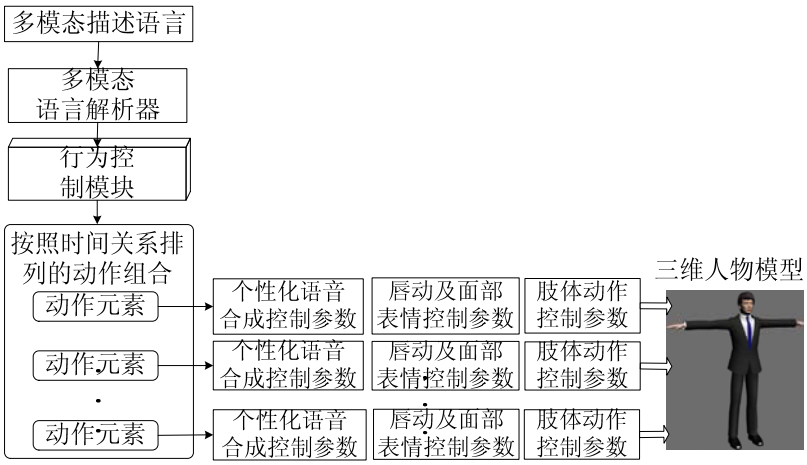

图 6 本文采用的虚拟人协同控制框架

4.2 语音同步下的人脸动画

由于人脸动画（尤其是唇动）与语音的关联程度较高，完全采用多模态协同置标语言的方式，难以实现人脸表情和唇动与语音的精确同步，也难以让表情和唇动反映出不同的语气变化或者语音中包含的情感变化。一般情况下解决这一问题的思路，是实现一种语音和唇动（或表情）的映射模型，常用的方法有：指数控制函数方法、人工神经网络方法、自组织映射网络方法、矢量量化方法和高斯混合模型方法等，这些方面通常只能针对语音与唇动之间的映射进行建模，难以实现基于语音的唇动和表情同时驱动。在这一点上，本文采用了基于 Fused Hidden Markov Model Inversion (Fused HMM Inversion)的语音到唇动和人脸表情的映射方法[26]。与传统的模型相比，该方法充分考虑了语音和脸部表情两个弱同步通道的特点，具有以下的优点：（1）语音和脸部表情可以单独训练成两个 HMM 模型，（2）通过 Fused HMM 的隐状态，采用最大熵准则和互信息最大准则，两个单独的 HMM 可以在结构上完整的链接在一起，从而在时间序列上将语音与脸部表情（包括唇动）这两个弱同步的通道有效的关联起来；（3）训练过程简单，在模型的复杂性和性能方面取得比较好的平衡。较之传统的单一基于语音的单一唇动或者人脸表情驱动方法，Fused HMM Inversion 方法，使虚拟人的唇部动画和脸部动话同步紧密结合，有效的改善了语音驱动下的人脸动画的表现力。

5 实验

5.1 多模态数字虚拟人表达

我们基于 Cal3D 平台软件[27]在主频为 2.6G，内存为 2G 的普通个人计算机上构建了一个数字虚拟人（如图 7 所示）。Cal3D 平台能很好的支持骨骼动画，所以 4.1 节记录的不同情绪的动作表达可以很好的迁移到数字虚拟人上，同时借助 4.2 的语音同步下的人脸动画、文本到语音合成技术、多模态协同控制语言，数字虚拟人能很好的表达不同情感强度下的姿态、表情动画。图 7 展示了虚拟人在疑惑的表情下，说出“请问您是要去哪个地方呢？”的其中某几帧的情况。

图 7 虚拟人在疑惑的表情下，说出“请问您是要去哪个地方呢？”表达序列中的某几帧

5.2 交通路况信息查询多模态人机对话系统

基于上述给出的多模态自然人机交互对话管理框架，本文设计了一个面向实用的多模态城市交通信息人机对话系统，其运行界面如图 7 所示，其中图 7 中的界面①-⑤分别了图

1 中系统结构对应的各模块，界面⑥对应了交通信息的中实时地图显示。

图 7 基于多模态融合的城市交通信息查询系统（数字分别对应图 1 的系统结构的模块）

该城市交通路况信息查询系统中，用户给出不同查询条件，针对用户的输入的不同查询信息、用户的不同姿态，以及用户针对实时路况信息反馈产生的情感变化，虚拟人会做出不同的反应和答复。同时，在回答和询问用户时，虚拟人的表情和动作也随着对话进程的不同而变化。图 8(a)(b)(c)列出了针对一个用户发出的路况查询请求，人机对话系统根据实时路况信息的不同，以及用户不同的语音、情感和姿态反馈，所可能产生的不同的对话过程。

a.

用户：到颐和园怎么走？（发起对话）
→**虚拟人**：请问您是要从哪里去颐和园呢？ （疑问的语调，身体略微前倾，表示征求意见）
用户：知春路
虚拟人：从知春路到颐和园，首先经过清华大学，然后路过圆明园，最后就到达颐和园了。（指向地图的实时路径显示，平和的语调，表示回答）
用户：那到鸟巢怎么去呢？
虚拟人：从知春路到鸟巢，首先经过北京航空航天大学，然后路过五道口，然后经过地质大学，最后就到达鸟巢了。（指向地图的实时路径显示，平和的语调，默认出发地为知春路）
用户：（查看系统给出的地图路径，一段时间没有说话）
虚拟人：目前北四环堵车现象严重（虚拟人主导对话，并给出实时交通信息）；
用户：（继续不说话，路出担心的表情）
虚拟人：建议换成地铁前往鸟巢，可以乘坐 10 号线，换乘 5 号线（平和的语气，语调上扬，给出建议）
用户：谢谢（用户对查询结果表示满意）
虚拟人：感谢您对我的服务感到满意，感谢您的查询（鞠躬，高兴的表情，并挥手做再见的姿势）

b.

用户：到颐和园怎么走？（发起对话）
→**虚拟人**：请问您是要从知春路去颐和园吗？ （疑问的语调，双手抬起，身体略微前倾，表示向用户确认信息）
→**用户**：（点头，不说话）
虚拟人：从知春路到颐和园，首先经过清华大学，然后路过圆明园，最后就到达颐和园了。（指向地图的实时路径显示，平和的语调）
用户：那到前门怎么去呢？
虚拟人：从知春路到前门，首先经过西直门，然后路过西单，最后就到达鸟巢了。（指向地图的实时路径显示，平和的语调）
用户：（满意的表情）
虚拟人：感谢您的查询（鞠躬，高兴的表情，并挥手做再见的姿势）

c.

用户：到颐和园怎么走？（发起对话）

→**虚拟人**：请问您是要从知春路去颐和园吗？（疑问的语调，双手抬起，身体略微前倾，表示向用户确认信息）

→**用户**：（摇头）

虚拟人：那您是想从哪里去颐和园呢？（疑问的语调，疑惑的表情，身体略微前倾，表示向用户征求意见）

用户：动物园

虚拟人：从动物园到颐和园，首先经过国家图书馆。。。（指向地图的实时路径显示，平和的语调）

用户：啊，不，从鸟巢去颐和园。（用户发现错误，改正出发地点，打断虚拟人说话）

虚拟人：从鸟巢到颐和园，首先经过中国地质大学，然后路过五道口，然后路过清华大学，最后就到达颐和园了。（继续指向地图的实时路径显示，平和的语调）

用户：（满意的表情）

虚拟人：下午 5 点后，会有雷阵雨，请注意出行安全。（继续指向地图的实时路径显示，人主导对话，并给出实时交通信息）

用户：谢谢（用户对系统的提示刚到满意）

虚拟人：感谢您对我的服务感到满意，感谢您的查询（鞠躬，高兴的表情，并挥手做再见的姿势）

图 8　针对一个用户发起的路况信息查询，系统可能出现的不同对话过程

5.3 主观测评

目前，针对多模态人机对话的自然度表达，国内外尚缺乏有效的评价标准。这里采用了主观评测的方法对本研究的多模态人机对话系统进行评估，由 20 名评测者参与了对北京市交通路况多模态信息查询系统的体验评估。这 20 名评测这在使用本研究的提出多模态人机对话系统，针对下列三种模式进行比较：1) 纯语音的交互方式，这种情况下，用户只输入语音，系统采用没有虚拟人的扬声器输出合成语音；2) 用户采用多模态的方式与系统进行交互，但是系统不处理用户的情感信息，同时在输出端，数字虚拟人采用随机的方式处理表情和姿态动画输出；3）用户采用本文提出的多模态自然人机交互方式和多模态情感表达虚拟人进行交互。评测者将按照以下标准对这些动画给出平均意见得分（mean of score，简称 MOS）。

5：自然，表现得像自然人与人对话的过程一样

4：比较自然，表现接近自然人与人的对话，但是不完美

3：中等，表现一般，对话形式略显呆板

2：不自然，自然人机对话体验较差

1：完全不自然

20 个评分者给出的分数经过平均以后的结果如图 9 所示（三组体验得分分别是 2.2 (Session (1)), 2.8 (session (2)), 3.8 (session (3))）。图 9 中，第一组纯语音的交互方式的平均 MOS 得分是 2.2，第二组交互方式平均得分是 2.8，本文的多模态自然人机交互方式的平均得分是 3.8，可以看到，前两种方法生成的人机对话模式，表现得不自然。由本文所提出包含用户情感行为、姿态变化的多模态自然人机交互方式能使人机交互的体验更为流畅，使得虚拟人与用户的交互显得比较自然活泼，明显地提高了人机对话的自然度和数字虚拟人的表现力，提高了用户的交互体验。

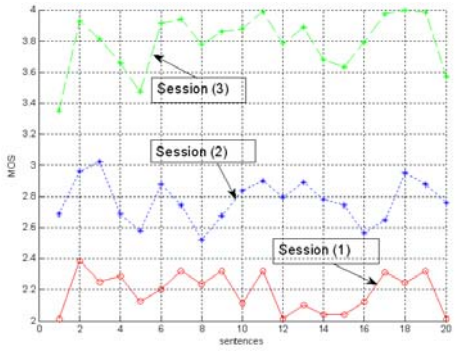

图 9 用户在不同交互模式下的 MOS 得分情况

6 总结

本文介绍了一种以语音为主，包含其他多种通道特征的多模态人机对话和表现策略，思路是以语音为基础，根据其他特征与语音的不同约束关系和时序关系，在对话主导和对话控制策略上进行融合反馈。本文详细的分析了多模态对话系统的实现方式，以及不同通道信息在对话过程中的融合方法，同时针对情感在对话系统中的作用进行了较为详细的讨论。在对话管理策略上，采用了基于事件驱动的混合主导的方式实现对话管理，并采用情感状态预测网络对用户的情感变化进行及时反馈。针对数字虚拟人的动作控制，提出了一种简化的多模态协同置标语言，分别在高层参数（如动作、姿态）和低层参数（如语音表现力、唇型、头姿）上实现了虚拟人的多模态展现，提高了虚拟人的表现力。应用于人机城市交通路况信息查询系统的实验表明，相对于简单问答的对话系统，本文提出的策略有助于提高人机对话的自然性。通过在语音、头姿、手势和情感等多通道信号进行融合处理的基础上构建的多模态对话系统，能够使用户较为灵活的向计算机提供信息；同时，借助于多模态协同置标语言(ML)，能够使得数字虚拟人在与用户交互时，也表现得更生动和更具有表现力，从而使用户在整个对话过程获得更为自然的体验。

参考文献

[1] http://en.wikipedia.org/wiki/Ananova

[2] http://www.mmdagent.jp/

[3] Fabrizio Morbini, David DeVault, Kenji Sagae, Jillian Gerten, Angela Nazarian, David Traum. FLoReS: A Forward Looking, Reward Seeking, Dialogue Manager In 4th International Workshop on Spoken Dialog Systems, 2012.

[4] D. Bohus and A. Rudnicky. Sorry, i didn't catch that! - an investigation of non-understanding errors and recovery strategies. In Proceedings of SIGdial, Lisbon, Portugal. 2005.

[5] David Goddeau, Helen Meng, Joe Poliforni,Stephanie Seneff,and Senis Busayapongchait, A Form-Based

Dialogue Management For Spoken Language Applications. International Conference on Spoken Language Processing(ICSLP'1996). Pittsburgh, PA, 701-704, Oct. 1996.

[6] Michael F, Mctear, Spoken Dialogue Technology: Enabling the Conversational User Interface，ACM Computing Surveys, 2002.

[7] Badler N, Steedman M, Achorn B, Bechet T, Douville B, Prevost S, Cassell J, Pelachaud C and Stone M Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversation agents, Proceedings of SIGGRAPH, pp 73-80, 1994.

[8] Pietquin, O.: A probabilistic framework for dialog simulation and optimal strategy learning. IEEE Transactions on Audio, Speech, and Language Processing, 589-599, 2004.

[9] Stefan Schwarzlery, Stefan Maiery, Joachim Schenk, Frank Wallhoff, Gerhard Rigoll, Using graphical models for mixed-initiative dialog management systems with realtime Policies, Annual Conference of the International Speech Communication Association - INTERSPEECH , pp. 260-263, 2009.

[10] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, Steve Young, A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies, Journal: Knowledge Engineering Review - KER , vol. 21, no. 2, pp. 97-126, 2006.

[11] Williams, J.D, Poupart, P. Young, P. "Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management", Proceedings of the 6th SigDial Workshop on Discourse and Dialogue, Lisbon, 2005.

[12] Steve Young, Using POMDPs for Dialog Management, Conference: IEEE Workshop on Spoken Language Technology - SLT , 2006.

[13] Antoine Raux, Maxine Eskenazi, A Finite-State Turn-Taking Model for Spoken Dialog Systems, Conference: North American Chapter of the Association for Computational Linguistics - NAACL , pp. 629-637, 2009.

[14] Chiori Hori, Kiyonori Ohtake, Teruhisa Misu, Hideki Kashioka, Satoshi Nakamura, Weighted Finite State Transducer Based Statistical Dialog Management, Conference: IEEE Workshop on Automatic Speech Recognition and Understanding - ASRU , 2009.

[15] Gokhan Tur Asli Celikyilmaz Dilek Hakkani-Tur, Latent Semantic Modeliong for Slot Filling In Conversationl Understadjing, 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing. pg. 8307. Vancouver, Canada, 2013.

[16] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, Steve Young, A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies, The Knowledge Engineering Review, Volume 21 Issue 2, Cambridge University Press New York, NY, USA, June 2006.

[17] F. Eyben, M. Wollmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, "On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues", Journal on Multimodal User Interfaces (JMUI) Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots, Vol. 3, No. 1-2, 7-12, 2010.

[18] Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee. Recent Approaches to Dialog Management for Spoken Dialog Systems . Journal of Computing Science and Engineering, Vol. 4, No. 1, Pages 1-22, March 2010.

[19] Carolis B.D, Pelachaud C, Poggi I, de Rosis F (2001) Behavior planning for a reflexive agent. In: Proceedings of the International Joint Conference on Articial Intelligence (IJCAI'01), Seattle, 2001.

[20] Aleksandra Cerekovic, Tomislav Pejsa, Pandzic Igors RealActor: Character Animation and Multimodal Behavior Realization System. IVA: 486-487, 2009.

[21] Van Welbergen H, Reidsma D, Ruttkay Z.M and Zwiers Elckerlyc. A BML Realizer for continuous, multimodal interaction with a Virtual Human, Journal on Multimodal User Interfaces,3(4):271-284, ISSN 1783-7677, 2010.

[22] Kipp M, Heloir A, Gebhard P, Schroeder, Realizing Multimodal Behavior: Closing the gap between behavior planning and embodied agent presentation. In: Proceedings of the 10th International Conference on Intelligent Virtual Agents, Springer, 2010.

[23] Jianhua Tao, Kaihui Mu, Jianfeng Che, Ya Li, Zhengqi Wen, Shifeng Pan, Lixing Huang, Le Xin, "Audio-Visual Based Emotion Recognition with the Balance of Do-mi-nances", International Conference on Artificial Intel-li-gence (ICAI1010), pp 100-110, Oct. 2010

[24] http://www.w3.org/TR/emma/

[25] http://www.w3.org/TR/speech-synthesis11/

[26] Jianhua Tao, Le Xin, Panrong Yin, "Realistic Visual Speech Synthesis based on Hybrid Concatenation Meth-od", IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, No. 3, March 2009, pp 469-477

[27] http://home.gna.org/cal3d/

HHME: Harmonious Human Computer Environment

Minghao Yang, Jianhua Tao, Hao Li, Linlin Chao

National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences

mhyang@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, hli@nlpr.ia.ac.cn, llChao@nlpr.ia.ac.cn

Key words：multimodal fusion；human computer interaction,；dialog management.

Abstract：During the dialogue, people naturally use multimodal information, e.g., facial expressions and gestures, in addition to using spoken interaction, to support the content expression. The paper proposes a framework on how to efficiently fuse these multimodal information with human-computer dialog model and finally creates a multimodal human-computer dialog system. The paper classifies the fused methods into three modes, complementary, mixed and independent, according to their relations between speech channel and other channels. For the dialog framework, the paper proposed a multimodal dialog management model by com-bining finite state machine, slot filling method and mixed initiative method. The new module can flexibly process the multimodal information during the dialogue. The paper also proposes a Multimodal Markup Language (MML) to control the action of the virtual human for the dialog system. The MML can help to coordinate the complicated actions

among different channels for virtual human. Finally, based on above technologies, the paper has created a multimodal dialog system and used it for weather information retrieval service