汉语口语对话中姿态与语音信息关系初探*

李爱军[1]，张利刚[1,2]，李　洋[1,2]，孟昭鹏[2]，王　霞[3]
(1.中国社会科学院 语言所，北京 100732; 2.天津大学 计算机科学与技术学院，天津 300072;
3.诺基亚中国研究中心，北京 100013)

文　摘：信息交互方式多种多样，以语音和姿态的表达最为自然，因此提高人机交互能力就需了解交际过程中的这 2 种模态对信息表达之间的关系。该文介绍了语音与姿态关系的相关理论和产生模型，并以电视访谈节目中自然对话的视频和音频数据为研究对象，对汉语普通话语音和姿态信息在交际过程中的关系进行了初步的研究。在语音学和姿态标注的基础上，分析了口语对话中焦点重音与姿态动作之间的关系，以及韵律边界和姿态边界之间的关系，研究发现语音上重音表达往往伴随较强烈的手部动作，而且此时手和头部动作之间有互补的现象；韵律边界和姿态边界没有时间上的对应关系，但有很大的相关性，这些结果都支持语音与姿态表达之间的关联理论。

关键词：自然口语；姿态；语音；多模态
中图分类号： H 017; TP 3

人机交互（HCI）的特征就是计算机或者智能机器人能够理解人的丰富多彩的言语行为，具有认知能力、甚至具有人一样的社会属性。因此人机交互是一个涉及多个学科的非常复杂的过程，通常需要设计一些问题、使用一些数据和方法来挑战某个单一学科。如用多个模态的数据和方法，来挑战语音一个模态的数据和方法，从而产生一些有效的方法。

当前的口语对话系统（Spoken Dialog System）面临着一些挑战：一是如何将口语对话中的各种信息加以整合，使得计算机不仅能够理解具有语言学意义的声音成份，而且还能理解和处理那些副语言学（如我们的情绪和情感的流露）和非语言学成份（如一些非言语噪音或者背景噪音）；二是如何提高人机语音交互的鲁棒性，例如口语语音识别中的一些重复、不连贯、不合语法的语段，口音和噪音等都会降低语音识别性能，语音合成中也需要提高合成语音的自然度；三是多模态语音的应用问题，即如何在人机交互系统中利用语音以外的其他模态如手势、面部表情等信息，例如利用视频信息研究言语行为与交际中姿态、表情等的关系，从而提高系统的“认知”能力。使得未来的智能机器人跟人一样具有一些自主性和认知能力，能够跟人类进行自然的交流，就像人与人之间交流那样自然。

近 25 年以来国际上对 HCI 技术进行了不懈的探索，比如德国 SMARTKOM 项目就是多模态对话系统的典范之一[9]。

本研究将目标锁定在第三个问题，即研究交际过程中的言语特征和姿态表达的关系，具体来说就是研究姿态表达和语音表达的关系。

国内外对语音和姿态 2 种模态各自独立的研究非常多，在情感计算和 Talking Head 等研究中将两者结合起来进行研究的也很多。

中科院自动化所对姿态和情绪的关系进行了研究，[1]认为人的姿态一般伴随着交互过程而发生变化，它们表达着一些信息。例如手势的加强通常反映一种强调的心态，身体某一部位不停地摆动，则通常具有情绪紧张的倾向。相对于语音和人脸表情变化来说，姿态变化的规律性较难获取，但由于人的姿态变化会使表述更加生动，因而人们依然对其表示了强烈的关注。他们还设计了智能科普娱乐机器人系统以及语音聊天机器人，其中涉及口语对话的言语行为研究。

文[6]报告口语中只有 7% 的信息是通过语言交流传递的，而 93%的信息是非言语信息，这其中 38%是通过声带音，而 55%是通过面部表情等其他方式传达的。

清华大学计算机系对可视化语音合成进行了研究和开发[2][3]，他们研究了语音的发声与人脸视觉特征变化的关联性，在语音可懂度和自然度的基础上，进一步为其增加情感和风格等，以实现具有表现力的个性化的合成。

香港中文大学系统工程与工程管理学部蒙美玲（Helen Meng）教授，对 M3 (mobile, multi-modal 以及 multilingual) 计算进行研究，开发了多语言口语系统，模拟人与人交互的人机交互系统。

文[4]对说话人的头、手和眉毛的动作 (visual beats) 与凸显的产生和感知进行了研究，发现说话人可以用很

*基金项目：国家“八六三”高技术项目(2006AA01Z138):

多信息来表示哪个词是凸显的，例如用音高重音和手势、头与眼部的运动(visual beats）等音频视频信息。实验发现，目标字和 visual beats 之间有显著的相关性，发音人产生一个可视的 visual beats 的时候，对应位置听起来就更重一些；参与者看到说话人的 visual beats，会使他们感知到有 visual beats 词比没有 visual beats 的词更重。

还有的研究发现 F0 和左边的眉毛运动相关性很高；音高和手势没有关系，但是经常同步。

文[5]对姿态中的不连贯现象，即与语音中的无声停顿（EPs）和填声停顿（FPs）相关的姿态进行研究。他们将姿态的停顿对应分为简单保持 simple holds(SHs：手和手臂都没有动作)和加强保持 augmented holds（AHs：手或者手臂轻微动作）。通过分析语音段中 EPs 和 FPs 的时间和个数分布，以及 SHs 和 AHs 的时间和个数分布，发现 EPs 和 SHs、FPs 和 Ahs 之间非常一致，因此，SHs 可以对应看成是姿态的正常短语的间断，而 AHs 对应代表姿态的检索或者概念化。还发现动作和语音上的停顿不同步。姿态表达我们思维的成像成分，跟语音不同，但是意义上并不是不同，而是传达的机制不同（图像而不是语音，2 维或 3 维而不是 1 维），他们共享的是交际特征。此外，他们观察到数据支持语音和姿态在交际行为中一样重要，两个模态相互支持使得表达的意图更加清晰。

文[7]对 Catchment 特征模型进行了详细介绍，这个模型解决了两个问题：第一，多模态信息交互中音频和视频信息之间的沟通桥梁；第二，不同模态之间的信息是怎样融合的。与整体姿态识别不同，Catchment 特征模型使用特征分解的方法，使得交叉模态可在话语规划和概念化层次得到融合。

文[8]对多模态的人类话语的姿态和语音之间的关系进行了详细的论述，把姿态分为 manipulative 以及 semaphoric 两种。前者是有意图控制的姿态，因此手或者手臂的动作和表达的实体之间密切关联，后者是指任何使用固定模式的静态或者动态的手和手臂姿势的系统。但是，当前的 HCI 系统需要的是自然话语对应的姿态，而且一定和语音有密切的关系。

我们对口语对话的语音特性进行了一些研究，发现人在交际过程中大量使用语音之外的其他信息，如副语言学信息、非语言学信息、姿态和表情等。[17，18]

而从分析国内外的研究，发现国外已经开始把各种身姿（脸、手和身体等）和语音两个模态的关系结合起来研究建模，而国内还只是关注面部姿态和语音表达关系。所以本研究的目的是探索汉语口语交际过程中的言语特征和姿态表达的关系。具体来说，就是研究姿态表达和语音表达的关系，两种表达系统对我们交际中信息传递的贡献、姿态表达和话语焦点关系、姿态表达和韵律结构的关系、姿态表达和情感情绪表达的关系、姿态表达和话轮转换的关系等等。本文介绍焦点重音和韵律结构与手部和头部姿态表达之间的关系研究。

1 姿态与言语产生的理论与模型

说话的同时经常伴随着一些姿态，姿态与言语产生的关系如何，各种理论的观点不尽相同。有的理论不承认有这种信息处理的共现关系，还有一些理论则认为姿态只表达言语所能传递的信息。本文介绍 3 种典型的理论。

1)无关联影像理论(free imagery hypothesis)，这个理论认为姿态是由工作记忆区的影像中产生的。姿态的内容是基于长时记忆生成的，或者是他一些思维过程生成的。最重要的是，他们是先于语言产生的(prelinguisticlly)，即与语言的表达无关。

有人认为姿态产生于工作区的空间影像，[11]与之不同的观点认为人们表达的姿态产生过程也是语音的产生过程，[12]即概念化过程 ，[13]先于语言产生的信息被反馈到语言形式化模块中。两种观点的相同之处是，姿态是在语音学形式化过程进行之前产生的。因此，无关联影像假设认为，姿态对信息的编码是不受语言表达的信息影响的。

2)词汇语义理论(lexical semantics hypothesis)，这个理论认为姿态产生是伴有语音词汇项的语义决定的。例如， 文[14]指出词汇项包含一些图形化的语义特征，图形化的姿态就是根据语义特征产生的，从而使语义特征得到空间解释。换言之，图形化的姿势是在语音产生的计算阶段产生的，而且是以抽象的形式，从语义上有组织词汇中选择词汇的某些意项之后产生的。

图形化的姿态来源于某些词汇项的观点认为各种形式的谈话显然是姿态产生的根源，确切地说就是谈话的词汇义项。词汇语义假设认为姿态的表达不会对共现的言语中没有表达的信息进行编码和表达。

3)关联理论(interface hypothesis)认为姿态是言语思维和空间思维之间共同作用产生的。交互的(interface)表达就是空间-运动表达(spatio-motoric representation)，即动作信息和空间信息以动作来表示，这种空间-运动的组织目标就是言语。因此，按照这个假设，姿态不仅对空间-运动相关物的特性进行编码(非语言学的)，而且对语言学编码可能性相关的信息进行结构化。[15，16]

关联理论不是词汇语义理论和无关联影像理论的混合，即一些姿态产生是由词汇语义假设决定的而另外一些是由影像无关假设决定的。关联理论认为，看到的姿态是由语言学信息和与之相关的空间-运动特性共同作用的结果，而这个相关物是不能通过同时出现的语音来表达的。

图 1 是基于关联理论的语音和姿态产生模型。这个模型主要阐明了姿态表达的内容是如何确定的，语音和姿态之间的同步现象不属于这个模型的考察范围。

Levelts 的语音产生模型与概念层的规划处理以及语音的形式化处理有本质的区别，概念化装置将交际意图转变成命题表达，被称为“pre-verbal message,’’送入形式化装置。形式化装置根据 pre-verbal message 的概念属性来检索词汇项，同时对一个句子的句法、词法和音系进行置标(make-up)。

模型中，Levelt 的概念化装置被分为两部分，第一个部分为交际规划装置(Communication Planner)，用来产生交际意图（Communicative Intention），并完成 Levelt 提出的宏观规划（Macro-planning）相同的功能。即大概确定表达的信息、表达信息的各部分顺序以及选择适当的言语行为。此外，还确定表达模态或方式。把表达模态选择处理纳入到概念化装置（Conceptualizer）是 Ruiter [12]首次提出来的。第二部分叫信息产生装置(Message Generator)，与 Levelt 的微观计划（Micro-planning）相同(即考虑句子的交际目的以及篇章内容，将命题 proposition 口语形式化)

模型的主要特性是：

1)交际规划装置确定表达的模态(方式)，但不必决定每个模态的具体表达信息。

2)一个姿态的内容是由下面几个方面确定的：

a)交际规划装置产生的交际意图；

b)基于想象和真实空间的特征选择的动作规划（Action Schemata）；

c)形式化装置到信息产生装置的在线反馈。

这些因素共同确定姿态的内容，单独一个因素不能确定姿态的内容，就是说姿态的内容不完全是由对交际做出贡献的机制确定的，而是一个产生动作的更为通用的机制(Action Generator)产生的。

3)在信息产生装置与动作产生装置之间，以及形式化装置跟信息产生装置之间存在在线的双向信息交互。这样就允许通过语言形式化的可能再现地确定姿势的内容。

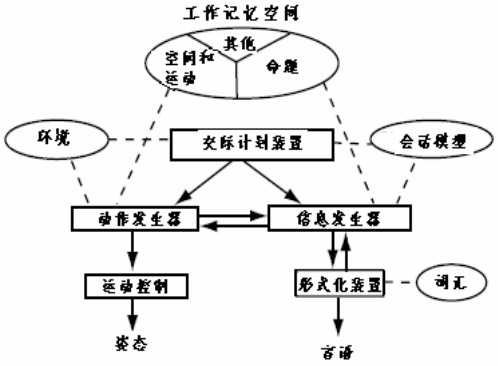

图1 基于关联理论的语音和姿态产生模型

本文赞同上述的关联理论以及基于这个理论提出的语音和姿态产生模型。本文以汉语自然口语对话为语料，研究姿态和言语之间的关系，考察姿态的表达与焦点和韵律结构之间的相关性，验证关联假设提出的语音和姿态产生模型。

2 语料

语料是用电视卡随机从电视台采集 4 段访谈节目的视频和同步的音频，选取的原则是讲话者有较为明显的手部或者头部动作。这 4 段视频文件分别为《采访马斌》(mabin: BTV-3，艺术大讲堂，时长 260s)，《名人堂》(mingren: BTV-5，名人堂，时长 160s)，《论人性》(people: 凤凰卫视，镲锵三人行，时长 74s)和《谈秦明》(qin: 河北卫视，文化大讲堂，时长 149s)。

3 多模态信息标注

项目中定义的多模态标注系统包含语音和姿态 2 个部分，其中语音的标注包括音段、韵律、副语言以及言语行为等语音和语言学信息标注；姿态的标注包括眼睛、眉毛、身体肩膀、手部、头部、面部表情和嘴巴。

3.1 语言学和语音学信息标注

语言学标注信息主要采用口语音段标注系统 SAMPA-C 和韵律标注系统 C-ToBI，此外还标注口语中的副语言学、非语言学的现象、以及话轮功能和话语功能。[17，18]音段标注主要标注音节和声韵母的边界。韵律标注主要标注重音和韵律结构。韵律边界包括三级边界，即主要韵律短语(语调短语 MAP)、次要韵律短语(MIP)以及韵律词(WORD)边界；对应重音(ST)也标注了两级，即语调短语重音“3”和次要韵律短语重音“2”。

口语中含有丰富的表达情感、情绪和态度的副语言信息，这里也进行了标注。

3.2 手部动作标注

本文的做法是把对立体空间的行为的标注简化为对一维空间的标注。采用感官上的强弱进行区分。所谓的强弱是一个相对的概念，是在一个有限的范围内进行比较得出的。范围界定是在一连串的动作或整个视频文件，按照动作的强烈程度用符号“0”、“1”、“2”、“3”进行标注。

定义 1

动作分割点(action division point)：人的手在做动作的时候往往会做出一组动作或单一的一个动作。单一的动作则直接以动作的开始和结束作为这个动作的分割点。一组动作中以动作的不连续点作为动作的分割点。动作的不连续是指动作的后一个状态和前一段时间比较发生了比较明显的变化或直接从动作当中感知出明显的停顿。比如，手的形状发生了变化，手的运动方向发生了变化(重复运动不包括在内)。

定义 2

视频单元(visual element)：以前后两个动作分割点为边界的一段视频称为视频单元。

如果整个手部成放松状态或放在桌上或腿上但手势不明显，标记为“0”。手部或手部加上胳膊做出某个姿式(与放松状态相比)持续一段时间，在这段时间内基本上保持静止状态或相当缓慢的出现运动状态，并且幅度也是很小，标记为“1”。手部或手部加上胳膊在一段时间内做出比较平稳的运动，而且幅度也不是很大，标记为“2”。手部或手部加上胳膊做出快速运动，或幅度很大，运动的速度很大，很强烈，标记为“3”。

3.3 头部动作标注

对头部的标注分类和含义如下：

■ 点头(nod)NOD

头的运动方向和肩膀垂直，比我们生活中感知到的点头的含义要广。一种是表示同意的点头，从上往下做一个弧线运动，然后收回到原始位置；另一种是运动到正前方，然后收回到原始位置。这两种情况都可以是往复运动，可以是一次，也可以是好几次重复的动作。

■ 摇头(shake)SHA

头的运动方向是水平的，成半圆形。可以从左肩移动到右肩，也可以沿相反方向移动。可以做一次或几次重复的运动，

■ 扭向一边或改变注视方向(side-turn)STN

多发生在旁边有另外的人的场合下，或私下里交谈，窃窃私语。为了使对方听清楚或自己听清楚对方的说话，头部所做出的动作。改变注视方向引起的头部的运动。可以在水平面上做半圆运动(但动作会比较慢，并且在一段时间内不会出现重复运动)，也可以在垂直面上偏向左边或右边。

■ 思考(thinking)THI

人在思考的时候头部容易做出的动作，低头一段时间或向上看一段时间。

■ 某些和社会关系相关的动作(social)SOC

一定和某些社会属性相关。比如学生挨罚的时候总是低着头，或者做错事的时候发生的动作。

■ 更复杂，待分类的动作(complex)COM

不属于以上几类的头部明显的动作暂时归到这一类。

多模态信息标注使用 ANVIL，[10]图 2 是标注示例。从上到下 1~4 层为姿态标注，5~13 层为语音和语言信息标注：汉字 HZ，拼音 PY，话轮 TURN，重音 ST，韵律词 WORD，次要韵律短语 MIP，主要韵律短语 MAP，副语言 MIS，话语功能层 FUNC。

图 2 标注示例

4 姿态表达方式与语音的关系

4.1 重音与手部动作的关系

对重音与手势之间的关系分析，考察重点是各级重音对应的手势，是否较重的重音对应较强幅度的手势，较弱的重音对应较弱幅度的手势？

定义 3：

相关：设 $T1$，$T2$，…，$Tn(n \leqslant 5) \in$ {HEAD,EYEBROW,HAND,BODY,ST}，如果 Ti 和 $Tj(i, j \in N, i, j \leqslant 5, i \neq j)$中的单元在时间轴上重合或部分重合，则这 2 个单元相关，记作($VALUE_{Ti}$，$VALUE_{Tj}$)。

本文从《采访马斌》段中提取了 174 个相关关系，《名人堂》段中提取了 84 个相关关系，《论人性》段中提取了 53 个相关关系，《谈秦明》段中提取了 88 个相关关系。

以 ST 为分类标准，得到了表 1 中的 HAND 和 ST 相关的个数和图 3 的相关个数分布图。

表 1 ST=3，2 和 0 时，手势级别(HAND=0~3)的分布

HAND 分类	相关个数			占总重音的比率%		
	ST =3	ST =2	ST =0	ST =3	ST =2	ST =0
3	121	63	24	59.3	48.10	37.50
2	34	42	25	16.7	32.10	39.10
1	22	17	15	10.8	13.00	23.40
0	27	9	24	13.2	6.80	

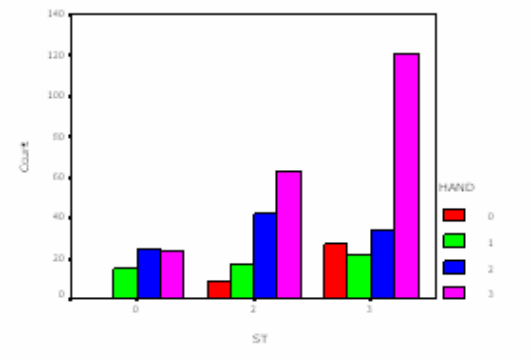

图 3 重音-手势相关个数分布

实验得到的概率矩阵分布如下：以所有手势对应的同一个 ST 值的总和为分母的概率矩阵 $R1_{ST\times HAND}$。其中 ST 依次取值为 3、2、0；HAND 依次为 3、2、1、0。

$$R1_{ST\times HAND}=\begin{bmatrix}59.3\% & 16.7\% & 10.8\% & 13.2\%\\ 48.1\% & 32.1\% & 13.0\% & 6.8\%\\ 37.5\% & 39.1\% & 23.4\% & 0\end{bmatrix}$$

以每类 HAND 个数为分母的概率矩阵 $R2_{HAND\times ST}$。其中 HAND 依次取值为 3，2，1，0，ST 依次取值为 3，2，0。

$$R2_{ST\times HAND}=\begin{bmatrix}58.2\% & 33.7\% & 40.7\% & 75.0\%\\ 30.3\% & 41.6\% & 31.5\% & 25.0\%\\ 11.5\% & 24.7\% & 27.8\% & 0\end{bmatrix}$$

Pearson Correlation 统计检验得出，重音级别与手势的幅度大小密切相关(P=0.03)，从 $R1_{ST\times HAND}$ 和 $R2_{ST\times HAND}$ 可以看出较高级别的重音对应较大幅度的手势比例较高，较低级别的重音对应幅度较大的手势比例较低。

这个结论支持文[4]的研究结果，也就是说话人用一定的空间运动信息来表达听觉上的语音凸显。而且实验的结果进一步说明了，动作的幅度大小与语音感知的凸显程度有很强的相关性。

4.2 头部动作与手部动作的关系

对定义 3 进行扩展，把相关性扩展到 3 层。所以 HAND，ST 和 HEAD 之间的相关性用($VALUE_{HAND}$，$VALUE_{ST}$，$VALUE_{HEAD}$)表示。如果只出现两层对应，那么第 3 个层级用 VALUE=0 表示。表 2 为当 ST=3 时，头部动作的个数随手势级别的分布如表 2 所示（前三行）。

表 2 ST=3 时，HEAD 个数分布

HAND 分类	个数	占总个数的百分比	3 相关个数
3	19	45.2%	121
2	8	19.0%	34
1&0	15	35.8%	49

可以这样假设，头部运动和手势对语音都会起到一定的辅助表达作用。现在把重音的个数按照手部动作的标注分成 3 类，把 HAND=1 和 HAND=0 归为一类，称为手势不明显组，HAND=2 为一类，HAND=3 为一类，这 2 类都是手势明显组。

从表 3 可以得到头部运动的个数在每组值中所占的百分比。结果如下：15.7% (19/121) < 23.5% (8/34) <30.6% (15/49)。即头部动作对重音的贡献比重随着手势级别的降低而增加，头部运动和手势对语音的辅助表达是互补的。

4.3 边界与手部动作的关系

数据提取方法：提取每一个动作分割点所对应的时间值，然后提取离这个动作分割点最近的左右两个韵律边界的时间点。计算左右两个时间点与动作分割点的差值，定义成 left 值和 right 值，如图 4 所示。这样，每一个动作分割点都会有一个 left 值和一个 right 值，然后再取 left 值和 right 值中较小的那一个，视其与动作分割点相对应，为一组。这样生成的两列数据就是本文要处理的数据。

图 5 是当韵律边界是语调短语(MAP)边界时对应的 left 值和 right 值模式图。图 6 是当韵律边界是次要韵律短语(MIP)边界时对应的 left 值和 right 值模式图。图 7 是当韵律边界是韵律词(WORD)边界时对应的 left 值和 right 值模式图。从图中可以看出，4 段视频的 left 值和 right 值的增长模式是一致的。并且通过单因素方差分析(ANOVA)检验表明，在 MAP 层，4 段视频的 right 值之间没有显著性差异(P>0.05)，4 段视频的 left 值数据中，仅《论人性》这一段和其他段之间有显著性差异(P<0.05)，其他各段之间没有显著性差异(P>0.05)。并且所有 left 值和所有 right 值之间没有显著性差异(P>0.05)。在 MIP 层和 WORD 层，各项检验均没有显著性差异(P>0.05)。

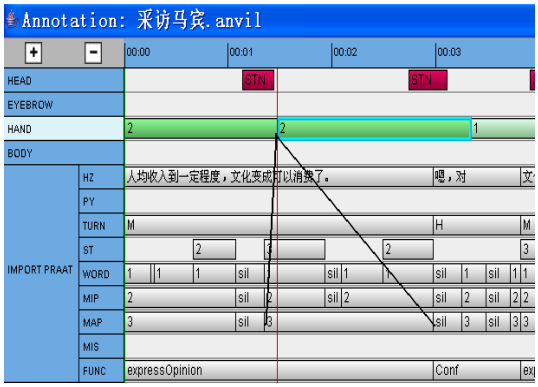

图 4 算法示意图

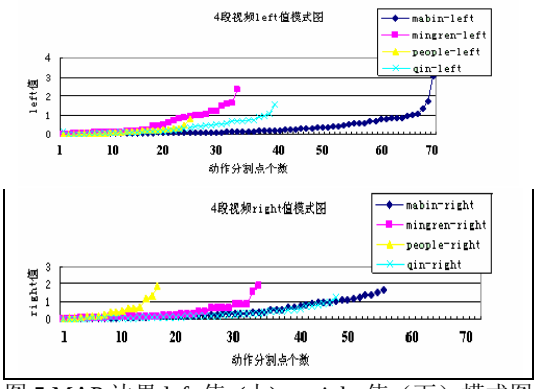

图 5 MAP 边界 left 值 (上)、 right 值（下）模式图

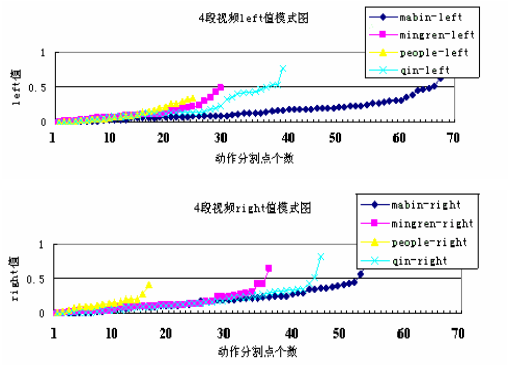

图 6 MIP 边界 left 值 (上)、 right 值（下）模式图

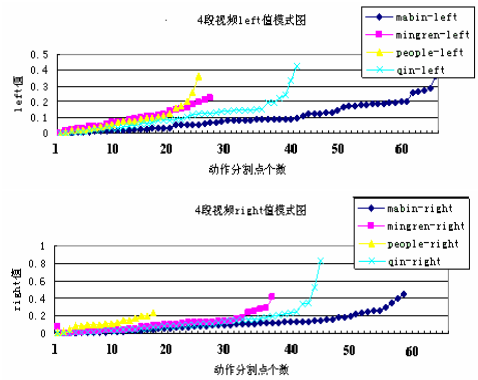

图 7 WORD 边界 left 值 (上)、 right 值（下）模式图

所以可以得出以下结论：

1、4 段视频在 MAP、MIP、WORD 与手势之间的关系表现出的模式是一致的。

2、标注系统的一致性得到了很好的验证。Pearson Correlation 统计检验表明：MAP，MIP，WORD 层的韵律边界与动作分割点的相关性很好(P=0.00)。

定义 4：

一致性：如果每组时间点的差值的绝对值小于等于所有 left 值和 right 值的平均值，则这组数据表现出了一致性。

MAP 边界中所有 left 值和 right 值均值为 0.391s。经计算得到所有的组中表现出一致性的有 213 组。语调短语层得到了 65.7%(213/324)的一致率。MIP 边界中所有 left 值和 right 值均值为 0.171s。次要韵律短语层得到了 61.1%(198/324)的一致率。WORD 边界中所有 left 值和 right 值均值为 0.111s。韵律词层得到了 56.9%(181/318，总个数的减少是由于语音标注的误差造成的)的一致率。

从分析数据看，韵律边界与手部姿势之间有很高的相关性。语调短语和次要语调短语边界与手部姿势之间的一致率较高，韵律词边界与手部姿势之间的一致率较低。这个结果验证了前面的语音与姿态产生的关联理论。姿态是由语言学信息和与之相关的空间-运动特性共同作用的结果。

动作分割点和韵律边界不是完全对应的，这也验证了动作和语音上停顿不完全同步，支持文[5]观点，手势编码对应的语音处理的单元是口语中的韵律短语(特别是语调短语)。

5 总结

本文提出了一套姿态标注系统，并对自然口语中的姿态进行了标注。

对姿态和韵律标注结果定量分析发现，姿态的强烈程度与语音的重音级别的相关性很好，较高级别的重音对应比较强烈的手势。证明了的说话人用一定的空间运动信息来表达听觉上的语音凸显的这个结论。[9]本文还发现在手势不明显的时候，头部动作会协助语音来表达信息，证明了头部动作和手势在表达信息时的互补性。

从边界和姿态的分析中我们得出，韵律边界与手部姿势之间表现出了很高的相关性，并且一致率随着韵律单元的变小而变小，证明了姿势是由语言学信息和与之相关的空间-运动特性共同作用的结果这个结论。动作分割点与韵律边界的不完全对应则恰恰说明动作和语音并不是同步发生的。本文的研究结论支持语音与姿态表达之间的关联理论。

本文还只是一个探索性的研究，我们所提出的一些算法和定义也有待改进。多模态标注系统也需要进一步细化和完善，并且在标注系统的实用性，一致性等方面需要进行量的考察。

*本文将发表于《清华大学学报》

参 考 文 献 (References)

[1] 陶建华、谭铁牛。数字化人类情感-和谐人机交互环境中的情感计算[J]，微电脑世界，2004 (1),29-32。
TAO Jianhua, Tan Tieniu. Emotion computing in harmonic HCI environment [J].*PC World China*, 2004 (1),29-32. （in Chinese）

[2] 王志明，蔡莲红. 动态视位模型及其参数估计[J]. 软件学报, 2003, 14(3): 461-466.
WANG Zhiming, CAI Lianghong. Dynamic Viseme and its pa

rameters estiming[J]. *Journal of software*, 2003, 14(3): 461-466. (in Chinese)

[3] 吴志勇，蔡莲红，蒙美玲. 可视语音合成中基于音视频关联模型的视位参数优化[J]，声学技术，2005, 24: 334-337.（第八届全国人机语音通讯学术会议论文集）
WU Yongzhi, CAI Lianhong, Hellen Meng. Viseme parameter optimization in AVCM for visualable speech synthesis[J]. *Technical Acoustics*, 2005, 24: 334-337. （proceedings of NCMMSC2005, in Chinese）.

[4] Emiel Krahmer, Marc Swerts. Hearing and Seeing Beats: The influence of visual beats on the production and perception [C]. // Proceedings of Speech Prosody. Dresden, 2006.

[5] Esposito A, McCullough K E, Quek F. Disfluencies in gesture: Gestural correlates to speech silent and filled pauses[C]// IEEE Workshop on Cues in Communication. Kauai, Hawaii, December 9, 2001. Also as VISLab Report: VISLab-01-18.

[6] Nick Campbell. Specifying Affect and Emotion for Expressive Speech Synthesis, Computational Linguistics and Intelligent Text Processing，Lecture Notes in Computer Science [M]，Springer Berlin / Heidelberg，2004.

[7] Quek F. The catchments feature model: a device for multimodal fusion and a bridge between signal and sense[J]. *EURASIP Journal of Applied Signal Processing*, Also as VISLab Report: VISLab-02-19,2004.

[8] Quek F, McNeill D, Bryll et al. Multimodal human discourse: gesture and speech [J]. *ACM Transactions on Computer-Human Interaction*, 2002,9(3):171-193.

[9] SmartKom: Foundations of Multimodal Dialogue Systems[M]. Springer, 2006.

[10] Michael Kipp. http://www.dfki.de/~kipp/anvil/.

[11] Krauss R M, Chen Y, Gottesman R F. Lexical gestures and lexical access: A process model . In D. McNeill (Ed.), Language and gesture [M]. Cambridge: Cambridge University Press, 2000:261–283.

[12] de Ruiter, J P. Gesture and speech production [D]. Nijmegen: Nijmegen University, 1998.

[13] Levelt, W J M. Speaking [M]. Cambridge, MA: MIT Press, 1989.

[14] Butterworth B, & Hadar U. Gesture, speech, and computational stages: A reply to McNeill [J]. Psychological Review, 1989,96: 168–174.

[15] Kita, S. How representational gestures help speaking. Language and gesture[M], Ed. by McNeill D,Cambridge: Cambridge University Press, 2000: 162–185.

[16] Kita S, özyürek A . What does cross-linguistic variation in semantic coordination of speech and gesture reveal? [J]. *Journal of Memory and Language*, 2003,48: 16–32.

[17] 殷治纲，汉语口语会话中的“嗯”、“啊”类话语标记研究[D]. 北京. 中国社会科学院研究生院，2007
YIN Zhigang. A study on “嗯/ng/”、“啊/a/”type of discourse markers [D]. Beijing: Chinese Academy of Social Sciences, 2007. （in Chinese）

[18] Aijun Li, Yiqing Zu. Corpus design and annotation for speech synthesis and recognition. in Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo(ED.), Advances in Chinese spoken language processing[M].World scientific publishing Co.Pte.Ltd, Singapore 2006: 243-268.

A Polite Study on the Relationship between Gesture and Speech Information in Chinese Spontaneous Speech

LI Aijun[1]，Zhang Ligang[2]，LI Yang[2], MENG Zhaopeng[2]， WANG Xia[3]

(1. Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China；2. Institute of Computer Science and Technology, Tianjin University, Tianjin, 300072, China;3. Nokia Research Center, China, No. 11, Hepingli Dongjie, Beijing 100013, China)

Abstract: Although there are various ways to communicate with each other for human beings, the most natural expression is speech and gesture. In order to improve the interactive capability for human computer interaction system (HCI), this paper makes a pilot study on the relationship between the two modalities of speech and gesture for Chinese spontaneous speech. Firstly, the paper introduces three relevant hypotheses and a speech and gesture production model. Secondly, a multimodal coding scheme is depicted and used for annotating four video and audio clips. Finally, the correlation between the speech stress and the hand gesture amplitude and the correlation between the gesture boundary and prosodic boundary are statistically analyzed. The results demonstrate that the stress expression usually accompanies with stronger hand gesture while the hand and head gestures are compensatory in this case; there is no time correspondence relation between prosodic boundary and gesture boundary, whereas they have significant correlation. All the results support the Interface Hypothesis.
Key words: Spontaneous speech, Gesture, Speech, Multimodal