语言文化背景与多模态情感表达和感知[1]

李爱军[1] 党建武[2,3] 方强[1]
[1]中国社科院语言所,北京
[2]天津大学 [3]日本北陆先端科技大学

题要
视频和音频两个模态的信息在交际过程中起着至关重要的作用，本研究将Brunswikian 透镜模型扩展到复杂传递通道，以及交际双方为不同文化背景的情况，考察不同模态下的情感表达（编码）和感知（解码），以及语言文化背景对情感表达和感知的影响。分析了 40 位中国听辨人（其中 20 位为日语学习者）以及 40 位日本听辨人 （其中 20 位为汉语学习者），对 2 位中国发音人和 2 位日本发音人在三种模态下（即只有面部表情、只有情感声音和面部表情与声音同时存在三种模态）7 种情感表达的感知模式，比较了不同模态下情感感知模式的异同，结合感知模式相似度的考察，发现发音人和听辨人的语言文化背景在不同模态下对情感感知的影响不同，只要有声音信息存在，语言文化背景的影响就存在，而对面部表情的感知，受到语言文化背景的影响较小。

关键词:情感表达，感知模式，多模态，Brunswikian透镜模型，语言文化背景

[1]本研究得到国家自然科学基金项目 NSFC No. 60975081 和 SPS Ronpaku 以及社科院创新工程项目支持。

1.引言

情感表达和交流是人类交际生活中的非常重要部分，对情感表达的研究可以追述到古人对七情六欲的研究、古希腊和古罗马人对修辞的研究。而近十年，在人机交互技术中，多模态情感计算更是成为研究的热点，其中涉及了情感识别、情感用户建模、情感结构建模以及虚拟人和机器人中的情感表达。但从理论上还缺乏广泛认同的情感理论基础以及对情感心理认知的研究。本研究涉及言语交际中，多模态情感表达和感知的关系，以及交际双方不同文化背景对情感表达和感知的影响。从理论上有助于揭示情感表达和感知的心理基础，从应用上来说，为多模态情感计算提供一些建模的依据。

最早进行不同文化情感研究的前辈有Charles Darwin(1988)，Ekman(1971)以及Izard(1994)，他们很早就注意到有不同文化背景的受试，对另外一种文化的演员的面部情感具有解码能力，提出如同面部表情一样，不同文化的人群对声音的情感表达也具有同样的解码能力，心理生态上情感具有普遍性的机制。

对不同语言文化的情感编码和解码的深入研究急需开展，以提供给语音技术所需的不同语言和文化背景的交互应用（Scherer, 2000)。

Scherer指出，研究不同文化和不同语言与情感语音的感知关系的理想途径就是在不同文化情况下对同样情感刺激进行评价，揭示出哪种相似的表达结构具有可比的评价和反应趋势。在跨文化研究中Abelin（2004)发现面部情感和声音情感识别的正确率比较，面部表情识别率远远大于声音通道的，多模态的情感表达（面部和声音一致的情感表达）在跨文化情感感知中效果更好。

Abelin和Allwood (2000)的研究发现，不同母语的人对情感的感知跟情感表达内容相关，特别是“Anger, Fear, Sadness以及 Surprise”这几种情感，而对其他情感如“Shyness, Dominance, Happiness和 Disgust”的感知，跟内容表达的关系不大，但他们没有给出合理的解释。他在另外一个跨文化的多模态实验研究中发现(2004)，跨文化的研究用多模态比声音一个模态更为成功。Yanushevskaya等（2008）发现每个文化的感知模式不同，对音高、嗓音音质以及音高+嗓音音质等参数的敏感程度在对不同情感具有不同表现。

Dang等（2009）对跨文化的情感因素进行分析，发现存在独立于各种语言和文化的语音情感感知因素，每一种情感即便是发音人按照自己的情感意图去说，也被感知成不同的成分，如果对某种情感的感知结果不好，那么一定会对另外一种情感有较高的感知结果，对三种语言背景的听辨人都有这样的表现，主成分分析法得到的结果显示，6种基本情感“Anger、Joy、Fear、Disgust、Surprise和Sad”可以近似地归为以“ Anger、Joy和Sad”为主要三种基本情绪，提出类似情感三色“调色板”的观点。

Scherer提倡使用Brunswikian透镜模型研究情感的编码、传递和解码过程，模型原理如图1所示。图中上半部分是整个模型的概念部分，说明了情感交际的整个过程。首先，发音人就情感状态进行表达(也叫编码)，表现在某些嗓音或者语言特征上，这些特征能从信号中客观测量到。具体来说，就是说话人的情绪唤醒伴随一定的生理变化，表现为呼吸、发声和调音等声学参数的变化。在模型中，这些声学特性与发音人的生态有效性（ecological validity）相关，从而使得话者对听者产生影响。从听者角度看，这些特征就称为远端特征（Distal cue）。远端特征可以通过各种传输通道传递到听话人的耳朵里，通过听觉系统进行感知。从感知者的角度看，模型中感知到的特征称为近端特征（Proximal cue）,因此近端特征就是感知者通过听觉系统感知到的特征，比如音高，响度等，在情感的感知中，还可以根据音乐感知特征，把近端特征分为声音响亮、顿挫、清晰、轻柔、洪亮等特征。模型中，话者表达的远端特征通过各种通道环境的传递之后，与感知者主观解码得到的近端特征不一定是对等的。从情感的感知来说，就是说话人所表达的情感，经过传输，感知人主观解码后感知到的情感不一定就是话者想表达的情感。

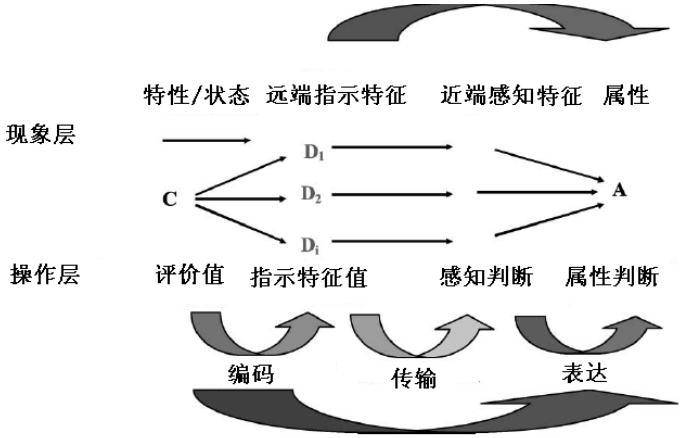

图1 改进的Brunswikian的透镜模型（Scherer 2003）

Brunswikia 透镜模型虽然主要针对声音模态提出来的，但我们可以将其扩展到多模态的研究中，比如应用到声音、面部表情、体姿等多个模态的编码、传递和解码的过程，考察不同模态的远端指示特征和近端感知特征的关系，甚至不同模态如视频和音频通道发生冲突时候的情感感知模式，也可扩展到话者和听者的跨文化的研究中。这也是我们这项研究对透镜模型的一个扩展。

我们采用 Brunswikian 透镜模型研究了跨文化的情感编码、传递和解码的一个案例(XXX,2010)，对一名中国发音人的 7 种情感（6 种基本情感+中性情感）进行感知实验，结果表明面部表情和声音情感信息在跨文化的情感解码中起着不

同的作用，面部信息的加入有益于情感的感知，听辨结果具有稳定性，跨文化的情感感知存在普遍的心理基础。

本文的研究进一步扩大了语料，发音人增加为 2 名中国人和 2 名日本人，情感听辨人为来自日本和中国的 80 位具有不同语言文化背景的学生（男女各一半），其中有 20 名不会日语的中国学生、20 名学习日语的中国学生、20 名不会汉语的日本学生以及 20 名学习汉语的日本学生。通过分析他们的感知模式，我们可以了解不同文化的情感表达差异，以及听辨人的语言文化背景在不同模态下对情感感知的影响，进一步解释 Brunswikian 透镜模型中，情感编码和解码与情感传递通道的关系，以及与表达者和听辨人语言文化背景的关系。

假设：某种语言的学习者的情感感知模式如果和非学习者有差异，并且与母语听辨人的结果更接近，那么就说明语言文化背景对情感感知模式有影响；否则，说明语言文化背景对情感感知模式没有明显的影响。

2.感知实验简介

2.1 实验材料

表 1 是汉语和日语的发音语料。10 个字面无情感趋向意义的短句分为两组（set1 和 set2）。汉语每组短句分别为 1-5 音节长，句法结构不同。汉语发音语料按照研究目的和语音语法等方面原则进行设计，首先是要考虑长度，1~5 个音节。目的是为在 McGurk 实验中，用一种情感的音频和另外一种视频的信息可以很好地搭配制作出所需的刺激（本文研究不涉及这个问题）;其次，考虑不同的句式和词性分配，字面含义为中性无情感含义；最后，语音上尽量覆盖三个顶点元音/a,i,u/。日语语料是对应的汉语翻译。

发音人共 4 名，1 男 1 女两名中国人 SONG 和 AN 为来自北京电影学院的学生，另外 1 男 1 女两日本学生 YD 和 ZHONG 为在中国的留学生，年龄都在 22-25 岁之间。

表 1 汉语和日语发音语料

语料说明	Set 1	Set 2
1~5 音节短句随机分配分为 2 组,句法结构不同	骂　ののしる	妈　お母さん
	踢球サッカーをする	大妈 おばさん
	吃拉面　ラーメンを食べる	奥运会　オリンピック
	足球比赛　サッカーの試合	打高尔夫　ゴルフをする
	滑雪场教练　スキー場のスキーコーチ	张雨吃拉面　張雨さんはラーメンを食べる

在语言所专业录音室用 Canon 数字录像机(Canon Power Shot TX1)对每个发音人的情感表达进行录音和录像，每一名发音人 10 个句子用 7 种情绪表达，正面情绪包括高兴(HAppy)，负面情绪包括难过(SAd)、生气(ANger)、厌恶(DIsgust)、害怕(FEar)。但惊讶(SUprise)以及中性(NEutral)情感不好按正负分类。发音人的情感表达采用激发的模式（王海波，李爱军， 2003）。

每个发音人的情感刺激制作为三类。（1）Aonly:只有声音（音频信息）的情感刺激。每组 7*5=35 个；（2）Vonly：只有面部表情没有情感声音的刺激。每组 7*5=35 个； （3）AVC: 面部表情和声音（音视频信息）同步一致的刺激。每组 7*5=35 个。4 位发音人三种刺激每组为 35*4*3=420 个，两组一共 840 个。

2.2 实验过程

共有四类听辨人（受试）：不会日语中国大学生(C)，会日语的中国大学生(CL)、不会汉语的日本大学生(J)和会汉语的日本大学生(JL)。每类听辨人为 20 人，由于听辨任务的时间太长会造成疲劳，我们将每类听辨人随机分为 2 组，每组 10 人，他们只参加一组语料的三个听辨试验。

中国的日语学习者（CL）均为北京理工大学日语系三年级本科学生，具有高级日语听说水平；日本的汉语学习者(JL)均为北京大学的日本留学生，学习汉语 3 年以上。因此，这些学习者的第二语言水平可以视为同样的级别，在感知实验中个体的差异可以忽略不计。不会日语中国大学生(C)为中国社会科学院研究生院语言学系学生，不会汉语的日本大学生(J)为日本北陆先端科技大学院大学的硕士和博士研究生。听辨人男女生人数各一半，听力正常。

80 名中日听辨人分别对 Aonly、 Vonly 以及 AVC 三种模态进行感知实验，表 2 是实验过程描述，每个听辨人的三个任务 T1~T3 至少相隔一周进行。听辨人除了进行情感类型判断，还进行情感特征判断（对特征的分析不是本研究关注的内容），并给出 1~5 分的打分，按照透镜模型和前人的研究，听辨人对情感和特征的判断设计为多选。图 2 为男发音人 SONG 的 Vonly 感知实验的界面，听辨人根据发音人面部表情，进行情感类型和强弱、面部动作特征和强弱的 5 度打分，可以多选，分数越高，表示强度越高。

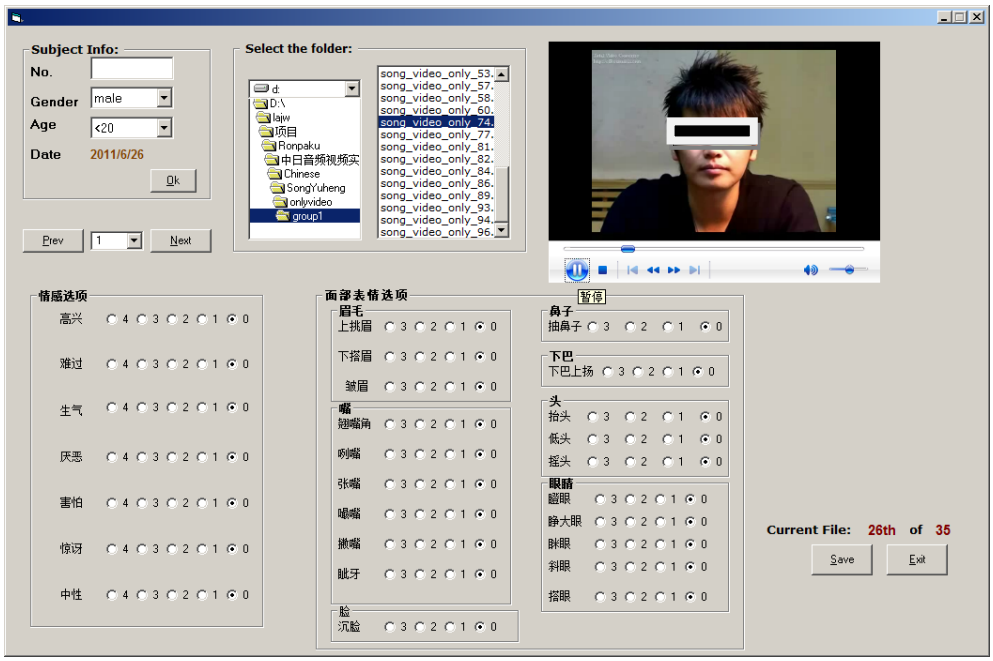

图 2 只有面部表情的 Vonly 模态听辨实验界面。右下部分选项为面部表情动作特征听辨选项，左下部分为表情评分选项。

表 2 感知实验描述

听辨实验编号	任务和过程
T1:Aonly	根据声音信息判断情感类型和强度，设为 5 级：0 最轻，4 最强，同时判断音色描写特征和表现强度，分为 4 级（0-3），情感和特征都可以多选。
T2: Vonly	根据视频信息判断情感类型和强度，也分为 5 级：0 最轻，4 最强，同时判断面部表情动作单元特征和表现强度，分为 4 级（0-3），情感和特征都可以多选。
T3: AVC	对原始的视频和音频信息一致的刺激进行情感类型和强度评价 5 级（0-4）、音色描写特征及其表达强度 4 级(0-3)，以及面部表情的动作特征及其强度 4 级(0-3)。听辨人对声音和面部表情特征的选择是自由的，有的可能只靠音色特征判断，有的靠视频判断，有的两者都起作用，由听辨人自主确定。

3.情感感知听辨结果分析

在透镜模型中，情感编码和解码涉及近端和远端特征，但本研究不涉及这两个特征分析，只关注实验 T1~T3 中对情感感知类型和强度的结果。通过分析解

码（感知到情感）后的感知模式，探讨经过不同情感传输通道，感知模式与话者和听者的语言文化背景的关系。

图 3 给出了感知模式受到语言文化背景影响的示意图。假设在某一个传输模态 M 下，话者用语言 A 的情感表达，如果被 4 类不同的语言文化背景的听辨人感知，即 A1 和 AB2、B1 和 BA2。其中 A1 和 AB2 的母语都为 A，但 AB2 还是语言 B 的学习者; B1 和 BA2 的母语都为 B，但是 BA2 还为语言 A 的学习者。感知模式如果出现语言学习者 BA2 的感知模式与母语听辨人 A1 和 AB2 的模式更接近的状态，就可以推断，在情感传输模态 M 下，感知模式受到听者和话者的语言文化背景的影响。

四类听辨人的感知模式的距离可以用相似度来测量，从而使图 3 的问题转化成计算四类听辨人的感知模式相似度和聚类分析问题。如果学习者 BA2 与 A1 以及 BA2 与 AB2 的相似度，和母语发音人听辨人 A1 和 AB2 的相似度都大于非母语 B1 与 A1 以及 B1 与 AB2 的相似度，那么就可以断定感知模式受到语言文化背景的影响。

本文的语言变量为汉语普通话和标准日语，模态 M 分别为 Aonly,Vonly 和 AVC。实验中，话者分别为 2 位日本人和 2 位中国人，通过 80 位不同语言文化背景的听辨人对 4 个话者情感表达的感知考察，对 3 个模态下情感表达的感知模式进行分析，探索感知模式在不同传输通道情况下，感知模式与话者和听者语言文化背景影响的关系。

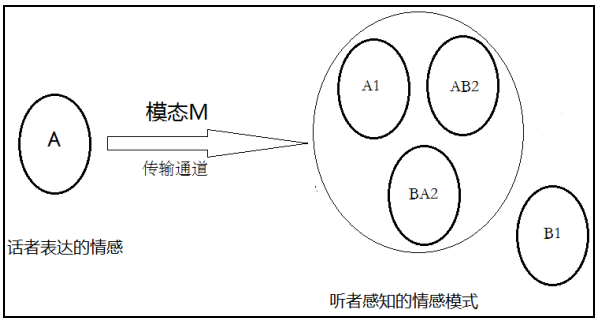

图 3 感知模式受到语言文化背景影响的示意图。话者情感表达经过通道为模态 M 传输后，被不同的听者感知。图上话者母语为 A, 有四类听者，他们的语言背景不同：A1 和 AB2 的母语背景也为 A, 但是 A1 不会 B, AB2 为语言的学习者；B1 和 BA2 的母语为 B, B1 不会说 A 语言，而 BA2 为 A 语言的学习者。

下面主要分析 M 为 Aonly, Vonly 和 AVC 三种模态下，4 类听辨人对 4 位发音人的情感类型和强度感知结果。首先分析感知模式，即发音人表达的某种情感，可能被听辨人感知为 7 种情感的强度平均得分。这里，根据透镜模型，实验设计

为听辨人根据自己的主观的判断，可以在 7 种情感中选择一种或者多种情感，7 种情感的平均得分分布表示了对应这种情感的感知模式（这里也称为感知向量）；其次分析感知模式之间的相似度，即比较中日听辨人之间，以及两组感知语料 Set1 与 Set2 之间感知向量的相似度，从而考察在三种不同的模态下，感知语料以及听辨人的语言文化背景是否对情感感知产生影响。

由于篇幅的限制，下面只列举 4 类听辨人听辨中国男发音人 SONG 和日本女发音人 YD 的听辨结果，但最后部分对所有的 4 位发音人的听辨结果进行总结和对比。

3.1 中国男发音人 SONG 的情感听辨结果

3.1.1 听辨结果的混淆模式

我们统计了 4 类听辨人对 7 种情感感知的混淆矩阵，为了便于呈现给读者，我们把听辨平均值用灰度图的形式展现给读者。图 4~图 7 分别表示了四类听辨人对中国男声发音人 SONG 在 Aonly、Vonly、AVC 三种模态下感知混淆模式图。图中纵坐标表示发音人表达的情感（intended emotions）。横坐标表示感知到的情感（perceived emotions）。每个小格子的颜色深浅代表感知的得分大小，颜色越深，表示感知到这种情感的得分越大。“An、Di、Fe、Ha、Ne、Sa、Su”分别表示“愤怒、厌恶、害怕、高兴、中性、难过和惊讶”。

以图 4 为例说明，图 4 为不会日语中国听辨人（C）对中国发音人 SONG 的情感表达，在声音 Aonly、Vonly 和 AVC 下 2 组语料（set1/set2）的听辨结果混淆模式图。每个方格的正对角线上的结果是 7 种情感正确感知的平均得分。比如我们看图 4 左上角的方格第一行，发音人的“愤怒”情感，对应感知最高得分的情感（颜色最深的）是“愤怒 An”，其次是“厌恶 Di”。

纵向比较图 4 和图 5 的三种模态的感知结果，可以看到中国听辨者（C/CL）对情感的感知在有面部表情信息加入后（中间和最下栏）明显好于只有声音的情感（上栏）。横向比较图 4 和图 5 的三个模态结果，没有明显的不同。接下来让我们纵向对比图 6 和图 7 日本听辨人（J/JL）的三种模态的感知结果，容易看到面部情感加入后（中间和最下栏），听辨结果好于只有情感声音的情况（上栏）。横向比较图 6 和图 7 的情况，还是有明显不同，比如图 6 和图 7 的第一栏的感知结果比较，图 7 在对角线上分布更集中，说明日本的汉语学习者对情感声音的感知结果更好。这一点预示着听辨人的语言文化背景对情感的感知会产生影响。

这些感知模式在三种模态下，不同听辨人之间是否有差异，即相似度如何呢？接下来将重点分析不同模态传递的情感，被不同语言文化背景的听辨人感知到的情感分布模式的相似度，从而考察不同模态下情感感知是否受到听辨人的语

言文化背景关系的影响

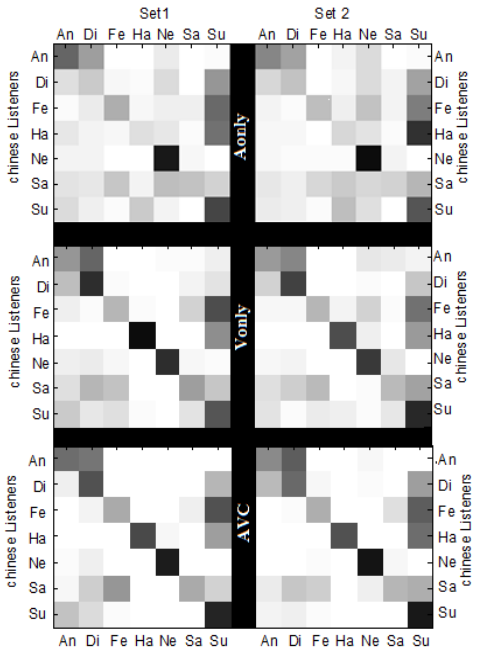

图 4 中国听辨者(C)在三种模态 Aonly（上）、Vonly（中）、AVC（下）对中国发音人 SONG 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

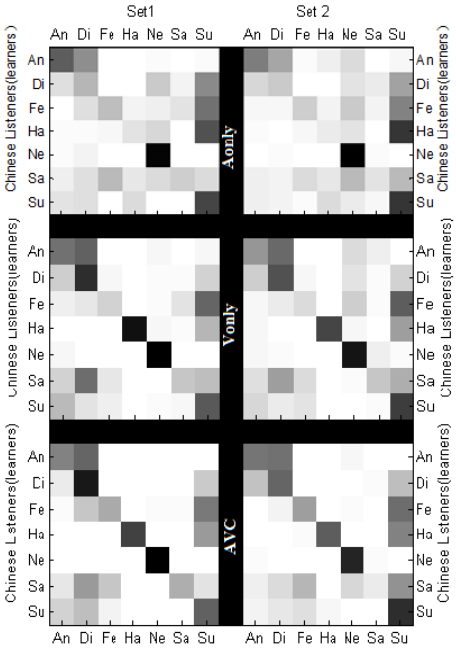

图 5 中国日语学习者(CL)在三种模态 A Aonly（上）、Vonly（中）、AVC（下）对中国发音人 SONG 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

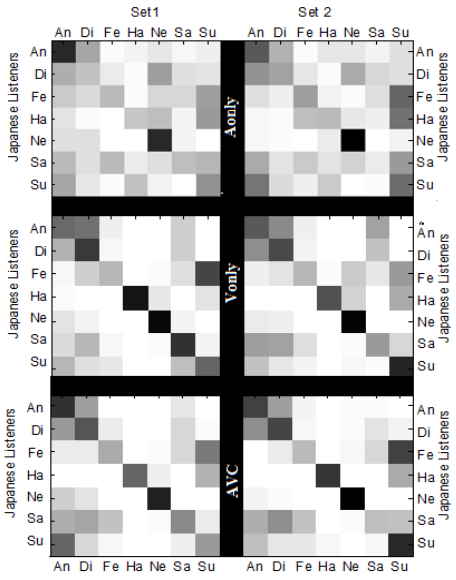

图 6 日本听辨者(J)在三种模态 Aonly（上）、Vonly

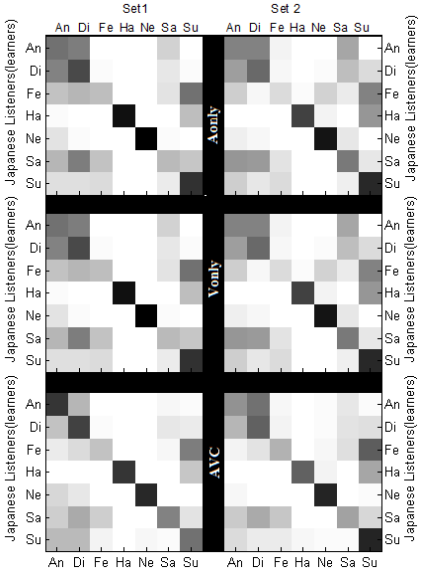

图 7 日本的汉语学习者(JL)在三种模态 AAonly

（中）、AVC（下）对中国发音人 SONG 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

（上）、Vonly（中）、AVC（下）对中国发音人 SONG 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

我们将采用相似度分析和聚类的方法，考察四类感知人的感知模式的差异（距离）。如果学习者的感知模式更接近母语感知者，则说明情感感知模式受到语言文化背景的影响。换言之，如果日本学习者和非学习者的感知模式与中国人的感知模式的差异不显著，则说明感知模式受到语言文化的影响较小。

相似度的度量方法采用 e-指数距离(Dang, et al. 2009): 两个感知向量 x_i, x_j 之间 e 指数距离定义为:

$$S(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\|x_i\| \cdot \|x_j\|}} \qquad (1)$$

表 3 是按照七种情感的感知得分，用公式（1）计算的四类听辨人之间在 7 种情感下的 e 指数距离 Ei（i=1…7）。D 是 7 种情感的距离矢量的模，$D = (\sum_{i=1}^{7} Ei^2)^{1/2}$， D 可以反映 7 种情感相似度的总体水平。

(1) 中国男发音人 SONG 的 Aonly 感知相似度分析

表 3 给出了 4 类听辨人对中国男发音人 Song 的 Aonly 感知相似度计算结果。第一栏是对第一组语料的感知相似度分析，第二栏是对第二组语料感知相似度分析，第三栏给出了同一类听辨人两组语料之间的相似度。对最后一列的 D^2 进行 K-means 聚类分析，如果可以得到日本的学习者和中国人的模式相似度，显著地大于不会汉语的日本人和中国人模式的相似度，即 C vs CL、C vs JL 与 CL vs. JL 分在一组，而 C vs J 与 CL vs. J 为另外一组，就可以断定学习者的模式受到语言文化的影响。

聚类结果用下划线来区分，具有下划线的是一类，没有下划线的为另外一类。同样，对 AVE.Set1 、AVE.Set2 和 AVE.Set1& 2 也进行 K-means 聚类分析。

(a) 第一组语料 set1：两组中国人之间模式分布相似度最大（D^2=2.426），会汉语的日本人情感分布模式与中国人的相似度（D^2=2.366, 2.322）跟中国人之间的相似度分在同一组（即 C vs. CLSet1、C vs JL Set1 与 CL vs. JL Set1 分为一组），显著地大于不会汉语的日本人跟中国人之间的相似度(P=0.02)。这个结果符合我们对语言文化背景影响感知结果的假设。

(b) 第二组语料 set2: 情况与第一组类似，聚类分析 C vs. CLSet2、C vs JL Set2 与 CL vs. JL Set2 分为一组，J vs. JL Set1、CL vs. J Set1 和 C vs. J Set1 分在另外

一组(P=0.029)。语言学习者对情感声音的感知模式与母语听辨人更相似，说明文化背景对于情感声音听辨有影响。

(c) 两组语料之间 Set1& 2 的感知模式，两组中国人之间的感知模式相似度显著高于两组日本人之间的相似度（P=0.014），说明母语听辨人对不同语料听辨的稳定性更好。

(d) 各类听辨人之间情感感知相似度的平均值(即 ave.set1/ ave.set2)聚类结果显示，各类听辨人之间对 7 种情感的感知相似度有差异。Set1 中，“中性、惊讶和生气”显著地高于其他 4 种情感（P=0.024）; Set2 中，“中性、惊讶和生气”显著地大于其他几种情感（P=0.04）。但是 AVE.Set1 、AVE.Set2 之间相关系数 R = 0.9354 ，也说明听辨语料对感知的影响不大。从 AVE.Set1& 2 结果看，4 类听辨人对 7 种情感听辨相似度分 2 组（P=0.05），“难过”和“害怕”受到语料内容的影响最大。

表 3 中国发音人 SONG 在 Aonly 模态下四类听辨者感知模式相似度比较

Song_Aonly	生气	厌恶	害怕	高兴	中性	难过	惊讶	D^2
J *vs.* JL Set1	0.926	0.639	0.774	0.763	0.943	0.704	0.782	2.108
CL *vs.* J Set1	0.913	0.696	0.782	0.759	0.913	0.773	0.774	2.130
C *vs.* J Set1	0.913	0.660	0.776	0.791	0.938	0.794	0.791	2.152
CL *vs.* JL Set1	0.859	0.801	0.906	0.897	0.957	0.746	0.953	**2.321**
C *vs* JL Set1	0.858	0.861	0.879	0.905	0.980	0.825	0.942	**2.366**
C *vs.* CLSet1	0.951	0.893	0.935	0.891	0.970	0.825	0.947	**2.426**
AVE.Set1	**0.90**	0.76	0.84	0.83	**0.95**	0.78	**0.86**	
CL *vs.* J Set2	0.898	0.684	0.823	0.808	0.963	0.729	0.797	2.168
C *vs.* J Set2	0.885	0.699	0.845	0.811	0.970	0.758	0.798	2.190
J *vs.* JL Set2	0.848	0.733	0.860	0.861	0.939	0.753	0.809	2.201
CL *vs.* JL Set2	0.885	0.813	0.805	0.886	0.919	0.810	0.939	**2.293**
C *vs.* JLSet2	0.896	0.837	0.811	0.894	0.933	0.818	0.917	**2.311**
C *vs* CL Set2	0.977	0.930	0.940	0.963	0.980	0.809	0.920	**2.468**
AVE.Set2	**0.90**	0.78	0.85	**0.87**	**0.95**	0.78	0.86	
J Set1 & 2	0.905	0.880	0.803	0.750	0.919	0.817	0.862	2.248
JL Set1 & 2	0.750	0.793	0.836	0.961	0.945	0.718	0.947	2.262
CL Set1 & 2	0.886	0.858	0.856	0.940	0.984	0.720	0.938	**2.346**
C Set1 & 2	0.914	0.920	0.869	0.855	0.977	0.768	0.941	**2.366**
AVE.Set1& 2	0.864	0.863	**0.841**	0.877	0.956	**0.756**	0.922	

(2) 中国男发音人 SONG 只有面部表情（视频信息）Vonly 模态下感知模式相似度分析

表 4 中国发音人 SONG 在 Vonly 模态下四类听辨者感知模式相似度比较

Song_Vonly	生气	厌恶	害怕	高兴	中性	难过	惊讶	D2
CL vs. J Set1	0.910	0.899	0.897	0.946	0.969	0.720	0.883	2.361
C vs. J Set1	0.882	0.927	0.918	0.905	0.948	0.752	0.903	2.362
C vs. JL Set1	0.885	0.900	0.838	0.951	0.933	0.828	0.922	2.368
J vs. JL Set1	0.961	0.916	0.872	0.946	0.981	0.745	0.866	2.384
CL vs. JL Set1	0.910	0.873	0.848	0.981	0.979	0.887	0.893	**2.411**
C vs. CL Set1	0.929	0.966	0.901	0.954	0.934	0.781	0.936	**2.424**
AVE.Set1	**0.913**	**0.914**	**0.879**	**0.947**	**0.957**	0.786	**0.901**	
C vs. J Set2	0.819	0.834	0.850	0.948	0.914	0.746	0.934	2.292
CL vs. J Set2	0.824	0.842	0.806	0.935	0.962	0.782	0.944	2.311
C vs. JL Set2	0.843	0.835	0.844	0.973	0.942	0.718	0.943	2.315
CL vs. JL Set2	0.840	0.847	0.855	0.976	0.984	0.745	0.955	2.354
C vs. CL Set2	0.920	0.936	0.896	0.970	0.943	0.813	0.940	**2.429**
J vs. JL Set2	0.937	0.920	0.814	0.933	0.960	0.926	0.956	**2.439**
AVE.Set2	0.864	0.869	0.844	**0.956**	**0.951**	0.788	**0.945**	
J Set1 & 2	0.923	0.938	0.780	0.877	0.982	0.789	0.858	2.331
JL Set1 & 2	0.916	0.895	0.795	0.914	0.960	0.842	0.973	**2.384**
CL Set1 & 2	0.904	0.928	0.857	0.909	0.959	0.844	0.905	**2.386**
C Set1 & 2	0.912	0.926	0.879	0.925	0.964	0.860	0.916	**2.414**
AVE.Set1& 2	**0.914**	**0.922**	0.828	**0.906**	**0.966**	0.834	**0.913**	

4 类听辨人对 SONG 的面部表情听辨结果的相似度分析列于表 4。总结如下：

(a) 对 set 1:中国人之间感知模式分布相似度(C vs. CL Set1)，和日本学习者同中国人之间的相似度（CL vs. JL Set1，C vs. JL Set1），大于不会汉语日本人同中国人的相似度（CL vs. J Set1 和 C vs. J Set1）。但聚类分析结果中国人和日本的汉语学习者并没有分在一组，也就是说两类日本人同中国人的相似度没有显著差异（P=0.004）。说明语言文化背景的影响有限，不显著。

(b) 对 set 2: 情况类似，聚类分析也是中国人之间和日本人之间分为一组，而中国人和日本的汉语学习者并没有分在一组，也就是说两类日本人同中国人的相似度没有显著差异（P=0.005）。说明语言文化背景的影响有限，不显著。

(c) 两组语料之间，中国人感知模式相似度略高于日本人，日本学习者跟中国分在一组，说明感知语料受到语言文化背景影响。（P=0.005）。与 Aoly 比较，学习者的模式更接近母语听辨人，所以影响更大。

(d) 对 AVE.Set1 和 AVE.Set2 的聚类分析（cluster=2），可以看到对 Vonly 的情感感知相似度有差异，第一组语料中，“难过”与其他的 6 种情感的感知相似度有显著差异（P=0.008），相似度最小；第二组语料中，高兴、惊讶和中性的相似度明显高于其他四类情感（P=0.004），但是两组语料的相似度大小顺序相当，相关性分析说明两组相关性高（R=0.852）。唤醒度低的负面情绪（难过和害怕），感

知的相似度小于其他的情绪。这一点与 Aonly 表现有区别。从 AVE.Set1& 2 也反映出来，语料对“害怕和难过”的影响更大一些（P=0.004）。

(3) 中国发音人 SONG 的音视频一致 AVC 模态下感知模式相似度分析

4 类听辨人对 SONG 的 AVC 情况下听辨结果的相似度分析列于表 5。总结如下：

(a) 对 Set1 以及 Set2: 中国人之间的感知模式最大，会汉语的日本人与中国人之间的感知模式，大于不会汉语的日本人与中国人之间的感知模式，在 cluster 设置为 2 的时候，会汉语的日本人都是跟某一组中国人的感知模式跟中国人分为一组（set1:P=0.004; set2: P=0.031）。因此，可以说此时语言文化背景对情感感知有一定的影响，但是影响程度小于 Aonly 模态。可以推断，情感表达内容对有声音的模态（Aonly, AVC）影响比面部表情(Vonly)要大。

(b) 对两组语料的 7 种情感相似度均值分析结果看，第一组语料 Set1 中，“生气、高兴和中性”的感知相似度显著地高于其他几种情感（P=0.012）；第二组中，“难过”显著低于其他几种情感（P=0.006）。AVE.Set1 和 AVE.Set2 之间的相关性 R=0.585。

(c) 表 5 第三栏表明“惊讶和难过”受到语料的影响显著大于其他几种情感（P=0.006）。

(d) 两组语料之间的感知结果分析按照听辨人的母语分类，听辨结果不能得到对应的 2 个分组（P=0.073）。说明大部分听辨人听辨感知模式与情感表达内容无关。

表 5 中国发音人 SONG 在 AVC 模态下四类听辨者感知模式相似度比较

Song_AVC	生气	厌恶	害怕	高兴	中性	难过	惊讶	D^2
CL VS. J Set1	0.868	0.847	0.877	0.915	0.925	0.853	0.788	2.299
C VS. J Set1	0.895	0.82	0.89	0.93	0.943	0.795	0.799	2.299
C VS. J LSet1	0.879	0.88	0.863	0.932	0.949	0.806	0.864	2.336
CL VS. JL Set1	0.85	0.904	0.897	0.937	0.924	0.878	0.931	**2.391**
J VS. JL Set1	0.962	0.92	0.885	0.908	0.983	0.923	0.825	**2.425**
C VS. CL Set1	0.965	0.918	0.887	0.981	0.961	0.811	0.899	**2.431**
AVE.Set1	**0.903**	0.882	0.883	**0.934**	**0.948**	0.844	0.851	
CL VS. J Set2	0.9	0.875	0.898	0.923	0.966	0.733	0.935	2.361
C VS. J Set2	0.863	0.853	0.949	0.914	0.98	0.796	0.93	2.381
CL VS. J LSet2	0.923	0.919	0.924	0.93	0.982	0.703	0.961	2.407
C VS. JL Set2	0.929	0.876	0.954	0.904	0.965	0.824	0.955	**2.425**
J VS. JL Set2	0.85	0.916	0.948	0.923	0.957	0.88	0.966	**2.436**
C VS. CL Set2	0.953	0.937	0.928	0.965	0.977	0.802	0.954	**2.467**
AVE.Set2	**0.903**	**0.896**	**0.934**	**0.927**	**0.971**	0.790	**0.950**	
CL Set1 & 2	0.971	0.881	0.903	0.946	0.956	0.683	0.883	2.364
JL Set1 & 2	0.819	0.937	0.895	0.904	0.948	0.911	0.85	2.371

J Set1 & 2	0.959	0.956	0.887	0.915	0.939	0.868	0.802	2.396
C Set1 & 2	0.944	0.897	0.964	0.935	0.973	0.806	0.94	**2.445**
AVE.Set1& 2	**0.923**	**0.918**	**0.912**	**0.925**	**0.954**	0.817		

3.1.2 中国发音人 SONG 在三种模态下的情感感知总结

比较图 4~图 7 的正确感知结果，可看到对于中国发音人，4 类听辨人的共同点都是面部表情的加入（Vonly 或 AVC）可以获得比只有情感声音更好的感知结果。比较表 3~表 5 第一栏和第二栏的 D^2，还可以进一步看到，对于非母语的听辨人，面部表情的加入后，7 种情感感知模式的相似度结果明显变大，说明感知效果有很大提高。

只有声音 Aonly 的时候，情感听辨受到语言文化背景的影响，汉语学习者与母语的感知模式更为接近；只有面部表情信息 Vonly 的时候，受到语言文化背景的影响小；面部和声音情感一致 AVC 情况下，感知结果虽然也受到文化背景的影响，但影响程度小于 Aonly 的情况。

3.2 日本女发音人 YD 的情感听辨结果

3.2.1 听辨结果的混淆模式

四类听辨人对日本女发音人 YD 的情感表达在三种模态下的感知结果见图 8~11。图 8 和图 9 是中国人的感知结果，都是 Vonly 和 AVC 的情况（第二和第三栏)下感知结果比 Aonly 好。图 10 和图 11 说明日本人的表情加入（Vonly/AVC）可以得到比情感声音 Aonly 更好的结果。中国的日语学习者的 Aonly（图 9 最上栏）比中国不会日语（图 8 最上栏）的感知结果要好，和日本人的 Aonly 模式（图 10 和图 11）更为相近。也预示了学习者情感的感知模式受到语言文化背景的影响。

因此，下面也通过相似度计算和分类统计，考察 3 种模态下听辨人语言文化背景跟感知模式的关系。

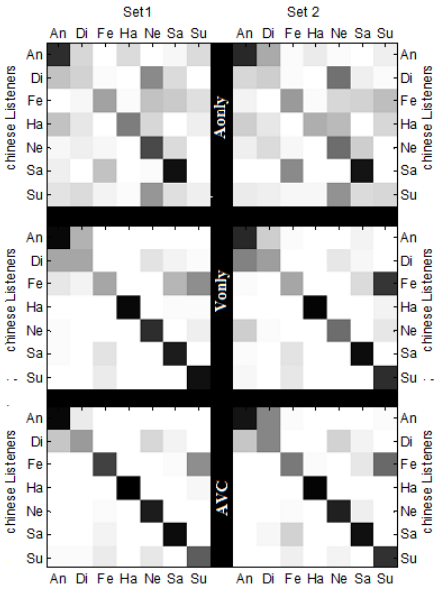

图 8 中国听辨者(C)在三种模态 Aonly（上）、Vonly（中）、AVC（下）对日本发音人 YD 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

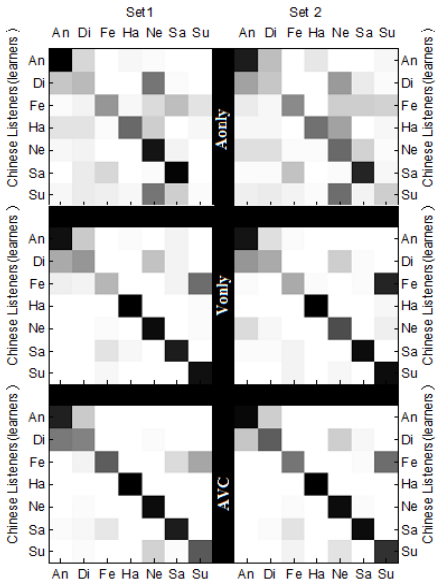

图 9 中国日语学习者(CJ)在三种模态 A Aonly（上）、Vonly（中）、AVC（下）对日本发音人 YD 的 7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

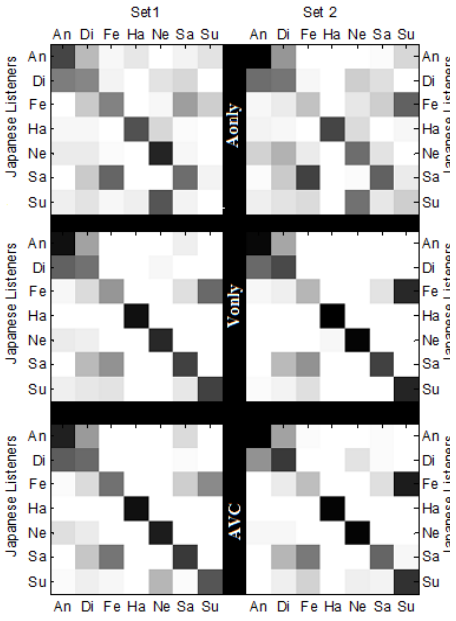

图 10 日本听辨者(J)在三种模态 Aonly(上)、Vonly（中）、AVC（下）对日本发音人 YD 的 7 种情感的

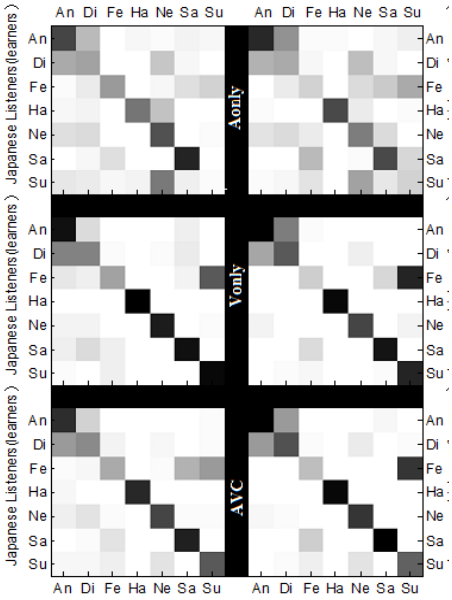

图 11 日本的汉语学习者(JL)在三种模态 Aonly（上）、Vonly（中）、AVC（下）对日本发音人 YD 的

感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

7 种情感的感知混淆模式图。左列为第一组语料 10 人感知结果，右列为第二组语料 10 人感知结果。

3.2.2 四类听辨人对日本女发音人 YD 的感知模式相似度分析

4 类听辨人在 Aonly、Vonly 和 AVC 三种模态下对日本发音人 YD 的 7 种情感的感知模式相似度在表 6~8 中给出。

(1) 日本女发音人 YD 的 Aonly 感知相似度分析

表 6 显示了两组语料在 4 类听辨人之间的感知模式相似度和聚类结果。

（a） 第一栏和第二栏的两组语料 set 1 和 set 2 的结果相似：会日语的中国人与日本人之间的相似度，大于不会日语的中国人与日本人之间的相似度，但是一半的学习者跟母语分在一组（J vs. JL、CL vs JL 为一组，显著地大于 C VS. JL，CL VS. J 、C VS. J。set1:P=0.032 ; set2:P=0.029）,结果表明情感感知模式部分受到语言文化背景的影响。

（b） 两组语料之间，两组中国人之间的感知模式相似度居中。从母语者听辨人的相似度判断，两组日本听辨人之间相似度一个为最小一个为最大，所以不能确定日本母语听辨人对不同的语料听辨结果一定好于非母语的中国人(P=0.082)。

（c） 从 7 种情感的平均相似度看，AVE.Set1 中,“生气、高兴、中性和难过”显著地高于“厌恶、害怕惊讶”（P=0.008）。 AVE.Set2 中, “生气、中性和难过”显著地高于“高兴、厌恶、害怕惊讶”（P=0.004）。 两组之间显著相关性（R=0.885**），说明语料对感知结果没有影响。

（d） AVE.Set1& 2 中，“生气、厌恶、高兴和难过”显著地高于“害怕、中性和惊讶”（p=0.012），说明“害怕、中性和惊讶”受到语料的影响更大一些。

表 6 中国发音人 YD 在 Aonly 模态下四类听辨者感知模式相似度比较

YD_Aonly	生气	厌恶	害怕	高兴	中性	难过	惊讶	D^2
C VS. J Set1	0.941	0.698	0.746	0.818	0.901	0.832	0.767	2.166
C VS. JL Set1	0.924	0.75	0.767	0.826	0.906	0.96	0.828	2.262
CL VS. J Set1	0.907	0.719	0.796	0.895	0.976	0.825	0.858	2.268
J VS. JL Set1	0.957	0.838	0.785	0.906	0.922	0.813	0.908	**2.322**
CL VS. JL Set1	0.918	0.778	0.823	0.917	0.91	0.977	0.884	**2.352**
C VS. CL Set1	0.92	0.874	0.884	0.89	0.907	0.96	0.835	**2.372**
AVE.Set1	**0.928**	0.776	0.800	**0.875**	**0.920**	**0.895**	0.847	
C VS. J Set2	0.92	0.696	0.734	0.696	0.855	0.884	0.821	2.131
CL VS. J Set2	0.919	*0.771*	0.704	0.845	0.869	0.83	0.91	2.218
C VS. JL Set2	0.951	0.743	0.764	0.679	0.932	0.916	0.85	2.220
CL VS. JL Set2	0.935	0.851	0.715	0.824	0.933	0.936	0.817	**2.281**
J VS. JL Set2	0.953	0.814	0.788	0.963	0.881	0.84	0.81	**2.293**
C VS. CL Set2	0.965	0.815	0.906	0.766	0.946	0.929	0.851	**2.342**

AVE.Set2	**0.941**	0.782	0.769	0.796	**0.903**	**0.889**	0.843	
JL Set1 & 2	0.931	0.948	0.699	0.866	0.855	0.905	0.748	2.261
C Set1 & 2	0.925	0.881	0.843	0.798	0.898	0.935	0.844	**2.318**
CL Set1 & 2	0.951	0.859	0.875	0.875	0.835	0.952	0.834	**2.339**
J Set1 & 2	0.905	0.945	0.746	0.966	0.846	0.954	0.858	**2.359**
AVE.Set1& 2	**0.928**	**0.908**	0.791	**0.876**	0.859	**0.937**	0.821	

(2) 日本女发音人 YD 的 Vonly 感知相似度分析

表 7 为 Vonly 时两组语料在 4 类听辨人之间的感知模式相似度聚类分析结果。

(a) 第一组语料中，C VS. JL Set 和 CL VS. JL 分在一组；而 CL VS. J Set 和 C VS. J 也没有显著差异（P=0.012），说明感知结果没有受到语言文化背景的影响；第二组语料中，J VS. JL 和 CL VS. JL 显著地大于 CL VS. J，分在两个不同的组里（P=0.043），说明有部分的影响。比较 Aonly，语言背景的影响还要小一些。

(b) 两组语料之间，中国人感知模式和日本人感知模式的相似度没有特别的分类规律 (P=0.116)，说明不同文化背景的感知人不同情感内容对感知结果没有影响。

(c) 7 种情感感知相似度的平均值看，第一组语料中，“生气、高兴、中性和惊讶”显著地高于“厌恶、害怕和难过”（P=0.006）；第二组语料中，“生气、害怕、高兴、难过和惊讶”显著地高于“厌恶和中性”（P=0.013）。两组语料的相关系数 R=0.605。

(d) AVE.Set1& 2 分类结果显示，“生气、高兴、难过和惊讶”受到语料的影响比“厌恶、害怕和中性”小。（ P=0.002）

表 7 中国发音人 YD 在 Vonly 模态下四类听辨者感知模式相似度比较

YD_Vonly	生气	厌恶	害怕	高兴	中性	难过	惊讶	D^2
C VS. J Set1	0.977	0.811	0.877	0.991	0.957	0.871	0.926	2.428
CL VS. J Set1	0.958	0.824	0.907	0.986	0.957	0.87	0.924	2.433
C VS. JL Set1	0.958	0.849	0.854	0.982	0.961	0.951	0.981	**2.474**
J VS. JL Set1	0.946	0.931	0.937	0.982	0.976	0.883	0.917	**2.486**
C VS. CL Set1	0.971	0.888	0.852	0.988	0.959	0.984	0.985	**2.508**
CL VS. JL Set1	0.981	0.842	0.933	0.992	0.976	0.947	0.979	**2.517**
AVE.Set1	**0.965**	0.858	0.893	**0.987**	**0.964**	0.918	**0.952**	
C VS. J Set2	0.936	0.856	0.965	0.984	0.817	0.871	0.974	2.425
CL VS. J Set2	0.937	0.799	0.968	0.99	0.884	0.863	0.962	2.426
C VS. JL Set2	0.907	0.845	0.943	0.986	0.882	0.985	0.968	2.467
CL VS. JL Set2	0.907	0.806	0.942	0.989	0.949	0.972	0.975	**2.477**
J VS. JL Set2	0.967	0.914	0.959	0.988	0.912	0.878	0.967	**2.491**
C VS. CL Set2	0.964	0.906	0.957	0.994	0.922	0.984	0.954	**2.526**
AVE.Set2	**0.936**	0.854	**0.956**	**0.989**	0.894	**0.926**	**0.967**	
C Set1 & 2	0.949	0.89	0.846	0.989	0.855	0.98	0.962	2.45
JL Set1 & 2	0.916	0.892	0.892	0.988	0.937	0.95	0.965	2.474
CL Set1 & 2	0.971	0.908	0.879	0.996	0.896	0.973	0.985	2.5

J Set1 & 2	0.982	0.947	0.899	0.982	0.947	0.996	0.945	2.533
AVE.Set1& 2	**0.955**	0.909	0.879	**0.989**	0.909	**0.975**	**0.964**	

(3) 日本女发音人 YD 的 AVC 感知相似度分析

AVC 两组语料和 4 类听辨人之间的感知模式相似度结果见表 8。

(a) Set1 中，中国学习者和日本人的模式，与学习者 CL 本人之间的模式正好分在两个不同的组（P=0.006），说明听辨结果受到语言文化背景的影响；Set2 中，尽管学习者与日本人的模式都大于非学习者 C 但 C VS. J 和 CL VS. J 分在两外一组，说明感知模式受到语言文化背景的影响较小(cluster=3 P=0.027)。

(b) 两组语料之间，中国人的相似度大于日本人的(P=0.025)。说明 AVC 情况下，母语者对不同语料的感知相似度不如非母语者们，有一定的语言文化背景的影响。

(c) 分析 7 种情感的平均值 AVE.Set1，感知的相似度有一定差异，“生气、高兴、中性、难过和惊讶”显著地大于“厌恶和害怕”(P=0.003); AVE.Set2：“生气、高兴和中性”显著地高于“厌恶、害怕、难过和惊讶”（P=0.06），但 AVE.Set1 和 AVE.Set2 之间有很高的相关性，说明尽管有一些差异，受语料的影响较小（R=0.905）。

(d) AVE.Set1& 2 的分析表明，“厌恶和害怕”受到语料影响稍大（P=0.008）。

表 8 中国发音人 YD 在 AVC 模态下四类听辨者感知模式相似度比较

YD_AVC	生气	厌恶	害怕	高兴	中性	难过	惊讶	D^2
C VS. J Set1	0.908	0.747	0.895	0.98	0.955	0.858	0.9	2.367
J VS. JL Set1	0.924	0.875	0.861	0.968	0.936	0.866	0.911	2.399
C VS. JL Set1	0.947	0.821	0.786	0.951	0.924	0.972	0.976	2.417
CL VS. JL Set1	0.98	0.92	0.824	0.954	0.914	0.988	0.94	**2.468**
CL VS. J Set1	0.935	0.939	0.926	0.985	0.956	0.866	0.935	**2.474**
C VS. CL Set1	0.951	0.778	0.924	0.991	0.98	0.973	0.933	**2.475**
AVE.Set1	**0.941**	0.847	0.869	**0.972**	**0.944**	**0.921**	**0.933**	
C VS. J Set2	0.966	0.83	0.866	0.997	0.961	0.845	0.949	2.43
CL VS. J Set2	0.96	0.9	0.859	0.998	0.983	0.831	0.941	2.451
C VS. JL Set2	0.973	0.847	0.873	0.995	0.958	0.978	0.903	**2.471**
CL VS. JL Set2	0.952	0.901	0.872	0.994	0.942	0.975	0.91	**2.477**
J VS. JL Set2	0.988	0.961	0.943	0.995	0.938	0.84	0.895	**2.483**
C VS. CL Set2	0.933	0.9	0.968	0.997	0.971	0.982	0.976	***2.544***
AVE.Set2	**0.962**	0.890	0.897	**0.996**	**0.959**	0.909	0.929	
JL Set1 & 2	0.925	0.89	0.781	0.96	0.949	0.96	0.964	2.435
J Set1 & 2	0.957	0.906	0.828	0.987	0.963	0.946	0.886	2.45
CL Set1 & 2	0.973	0.826	0.885	0.997	0.988	0.98	0.924	**2.489**
C Set1 & 2	0.904	0.923	0.906	0.992	0.978	0.966	0.915	**2.49**
AVE.Set1& 2	**0.940**	0.886	0.850	**0.984**	**0.970**	**0.963**	**0.922**	

3.2.3 日本发音人 YD 在三种模态下的情感感知总结

比较图 8~11 的感知结果，4 类听辨人都是在面部表情加入后，感知的正确率较声音情感高。在 Aonly 和 AVC 情况下，感知模式受到语言文化背景的影响，学日语的中国听辨人的感知模式更接近日本听辨人的。只有面部表情 Vonly 的时候，中国听辨人的感知模式受到语言背景的影响较小。两组语料之间的感知模式没有明显差异。

4.总结与讨论

4.1 四位发音人的情感感知模式总结

我们也对另外一名中国女发音人 AN 和一名日本男发音人 Zhong 的情感感知模式进行了聚类分析，得到了类似的结果。

在情感表达有面部视频信息加入后，所有文化感背景的听辨人的感知结果正确率得到提升。

在 Aonly 模态下，语言学习者的听辨结果分布模式与母语的分布模式更为接近，语言文化背景对情感听辨分布模式有明显影响或者较大的影响。在 Vonly 模态下，听辨人的语言文化背景对情感感知的影响没有影响或者影响较小。在 AVC 模态下，发音人的感知模式尽管受到听辨人的语言文化背景的影响，但是这种影响比 Aonly 小，介于 Aonly 和 Vonly 之间。

表 6 总结了两组语料对应 7 种情感相似度分组情况，以及对两组语料感知结果比较。三种模态下，四位发音人的情感感知相似度较高的情感各不相同，但是“生气、高兴和中性”情感感知的相似度均较高。四位发音人的 7 种情感中，负面情感“害怕、难过和厌恶”受到语料的影响比其他几种情感大，“惊讶”的表达可以为负面情感也可以是正面情感，发音人之间有差异，所以也体现了部分情感容易受到语料的影响。这一点与 Abelin 和 Allwood (2000)的研究发现部分相同，不同的是“厌恶和愤怒”，我们这里“愤怒”情感不容易受到语料内容影响，而厌恶反而容易受到发音内容的影响。比较特殊的是日本发音人 YD 的“中性”情感也受到了发音语料的影响。除了日本发音人 Zhong 的 Aonly 下的 7 种情感，两组语料相似度结果相关性很低（P=0.178）外，其他发音人的各个模态的情感表达在两组语料下的相似度模式相关性都很高。可以推测，可能是这位日本男发音人的情感声音表现力较弱造成的。

表 6 四位发音人感知结果总结

An	Aonly	Vonly	AVC
Set1 情感相似度分组	生高中难惊>厌害	生高中>厌害难惊	生高中>厌害难惊
Set2 情感相似度分组	生中>厌害高难惊	生厌高中>害难惊	生高中>厌害难惊
Set1&set2 影响大的情感	厌难	厌害难惊	厌难惊
R	0.586	0.972	0.973
Song	Aonly	Vonly	AVC
Set1 情感相似度分组	中惊生>厌害高难	生厌害高中惊>难	生高中>厌害难惊
Set2 情感相似度分组	中高生>厌害难惊	高中惊>生厌害难	生厌害高中惊>难过
Set1&set2 影响大的情感	害难	害难	惊难
R	0.935	0.852	0.585
YD	Aonly	Vonly	AVC
Set1 情感相似度分组	生高中难>厌害惊	生高中惊>厌害难	生高中难惊>厌害
Set2 情感相似度分组	生中难>厌害惊高	生害高难惊>厌中	生高中>厌害难惊
Set1&set2 影响大的情感	害中惊	厌害中	厌害
R	0.885	0.605	0.905
Zhong	Aonly	Vonly	AVC
Set1 情感相似度分组	生厌害高中难>惊	生高惊>厌害中难	生高中惊>厌害难
Set2 情感相似度分组	生害高中惊>厌难	生厌高中惊>害难	生高中>厌害难惊
Set1&set2 影响大的情感	厌害惊	害难	厌害难惊
R	0.178	0.823	0.944

4.2 语言文化背景与多模态情感表达与感知关系

通过上面对 2 位中国发音人和 2 位日本发音人的 7 种情感表达，80 位语言文化背景不同的听辨人的在三种模态下情感感知分析，我们总结了这 4 位发音的情感感知在三种模态下受到听辨人语言文化背景影响程度的分布特征。如表 7 所示， 表中‘++’表示在对应的模态下，发音人情感的感知受到听辨人语言文化背景的影响明显,‘+’表示有一定的影响,‘-’表示影响较小或者没有影响。

表 7 情感感知模式在不同模态情况下受到听辨人语言文化背景的影响总结

发音人	Aonly	Vonly	AVC
中国女发音人 AN	++	-	+
中国男发音人 Song	++	-	+
日本女发音人 YD	++	-	+
日本男发音人 Zhong	++	-	+

通过上面分析看到，发音人对情感的编码存在跨文化的共同基础，但也存在个体编码的差异（李爱军等，2008；Dang et.al., 2009），面部表情的加入有助于情感的识别；个人情感表达有差异，比如中国女发音人的‘难过’和男发音人的‘愤怒’的声音表达效果都好于面部表情，但总体来说“厌恶、难过和惊讶”等声音的感知结果都不是很好。

4.3 对 Brunswikian 透镜模型的贡献

日常情感交流是一种多模态的交际过程，人们通过文字、语言、面部表情和体姿等各种方式传情达意。

关于情感理论，影响较大的主要有三个(Scherer & Ellgrin 2007； Scherer 2003）:除了上面介绍的 Brunswikian 透镜模型，还有（1）离散情感理论，最早可以追溯到达尔文，认为情感是由离散的基本情绪构成的，是进化的结果并具有普遍性。存在一定数量的、进化上连续的基本情感，通过特定的条件触发，并通过相互关联的特定情感响应编程加以区分，其中响应编程是普遍的与生俱来的。目前比较认同的 6 种基本的情感为生气、高兴、害怕、难过、厌恶和惊讶。（2）成分情感模型-componential emotion models，认为面部情感表达的每个成分由评价结果确定并对运动行为产生影响。(Scherer, 2003； Grandjean & Schere , 2008)。

从言语交际过程看，Brunswikian 透镜模型更具有建模方面的优势，涉及了编码和解码端特征、传输通道特性等系统因素。Brunswikian 透镜模型最早只是应用到情感声音交流，我们将其扩展到声音、面部表情等多模态，考察情感的编码、传递和解码的关系，通过对情感感知模式的分析，了解不同模态情感表达、传递和解码的过程，以及交际双方语言文化背景对情感交流的影响。

要全面了解该模型的过程，涉及的研究非常多。比如尽端和远端特征的分析、发音人表达的异同、听辨人感知的异同等等问题，本文只是涉及了很小的一个部分，即多模态情感交流与交际双方语言文化背景的关系问题。

本研究的发现是：从解码和情感传递通道的关系看，在声音 Aonly 模态和声音与面部表情一致的 AVC 两个模态中，存在听辨人的文化背景对感知的影响，**也就是说只要情感传递通道有声音信息存在（音频或者音视频两种模态），情感的感知就会受到听辨人语言文化背景的影响**，具体来说，如果听辨人与发音人语言文化背景一样（都说母语）或者相近（听辨人为第二语言学习者），情感感知模式就更接近；**而对面部表情的感知，受到听辨人文化背景的影响比较小**。

本研究中的感知实验过程复杂，四分之一的实验是在日本完成的，整个实验几乎历时一年才完。目前发音人只有 4 位，今后有条件应该扩大发音人和情感语料，特别是增加一些自然口语情感的对比研究。美国听辨人的实验也在考虑中，这样将增加听辨人的文化差异，结果将更有意义。

参考文献

Abelin, A.; Allwood, J. 2000. Cross linguistic interpretation of emotional prosody, *Proc. ISCA Workshop on Speech and Emotion*, Belfast.

Abelin, Å. 2004. Cross-cultural multimodal interpretation of emotional expressions – an experimental study of Spanish and Swedish. In *SP-2004,* 647-650.

Dang, J. W., Li A. J., Erickson, D., Suemitsu, A., Akagi, M. Sakuraba, k., Minematsu, N. & Hirose, K. (2009). Common Factors in Emotion Perception Among Different Cultures. In *APSIPA ASC 2009,* 538-544.

Darwin, C. 1998 *The expression of the emotions in man and animals*. London: John Murray (reprinted with introduction, afterword, and commentary by P. Ekman, Ed.). New York: Oxford University Press. (Original work published 1872).

Ekman, P., Friesen, W. V. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124– 129.

Grandjean, D. & Scherer. K. R. 2008. Unpacking the Cognitive Architecture of Emotion Processes. *Emotion*, *8*, (3): 341–351.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 1994, 115:288–299.

LI，A. j., Shao, P. F. & Dang, J. W. 2009 .A Cross-cultural and multi-modal investigation on emotion expression (in Chinese). Journal of Tsinghua University (Science and Technology). Vol49.S1.

李爱军,邵鹏飞,党建武 2009. 情感表达的跨文化多模态感知研究. *清华大学学报*（自然科学版）49 卷,S1 期.

Scherer, K. R. & Ellgring, H. 2007. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal. *Emotion* ,7,(1): 113–130.

Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227-256.

Scherer, K. R. 2000. A Cross-Cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *ICSLP2000*.

Tomkins, S. S. 1984. Affect theory. In K. R. Scherer & P. Ekman (Eds.), Approaches to Emotion. Hillsdale, *NJ: Erlbaum*.1984:163–196.

王海波,李爱军（2003）. 普通话情绪语音库的建立及听辨实验.*第六届全国现代语音学学术会议论文集.*

Yanushevskaya, I., Chasaide A.N. & Gobl, C. 2008. Cross-Language study of vocal correlates of affective states. *Interspeech 2008*.

A CROSS-CULTURAL AND MULTI-MODAL INVESTIGATION ON EMOTION EXPRESSION AND PERCEPTION

Aijun LI[1], Qiang Fang[1], Jianwu Dang[2]

[1]*Chinese Academy of Social Sciences, Beijing*

[2] *Tianjin University, Tian Jin and Jaist, Kanazawa*

In the interactive communication, the encoding and decoding of information of various modalities and the comprehensive understanding take place all the time. This research concerns the process of encoding and decoding of facial and vocal emotional expression through a cross-cultural perceptual experiment from the perspective of psychology and cognition by using the framework of Brunswikian lens mode.

The stimuli were produced from four speakers (a female Chinese speaker, a male Chinese speaker, a female Japanese speaker and a male Japanese speaker) in three conditions of congruent Audio-video (congruent facial and vocal expression, AVC in short), Audio-only(vocal expression) and Video-only(facial expression). The text of the speech is 5 neutral phrases with 1 to 5 syllable length. The speakers expressed these phrases in 7 emotions including neutral states: happy, sad, surprise, angry, disgust, fear and neutral. The subjects are 40 Chinese and 40 Japanese including half language learners to exam how the culture background of language affect the expression and perception patterns in different transmitting modalities.

The perceptual results reveal that although individual speaker may employ different way to express emotions, common patterns still exit cross culture in encoding stage. From the decoding aspect, language culture will impose an influence on the identification of emotion. In AVC and A-only modalities, language culture backgrounds affect the perceptual patterns of emotions more than those in V-only state. In other word, as long as vocal modality exists, the decoding of the emotion has a relatively higher interaction with language culture background of the listeners, while it has smaller interaction with listener's language background for facial expression alone.

KEYWORDS

Brunswikian lens mode， Emotional expression， Multimodality， Language culture background， Emotion perception

此文发表在石锋、彭刚主编，《大江东去--王士元教授 80 岁贺寿文集》，香港城市大学出版社，2013。P306-332。ISBN 978-962-937-220-0.