文章编号:1004-9037(2014)02-0157-14

实用语音情感识别中的若干关键技术

赵 力 黄程韦

(东南大学信息科学与工程学院,南京,210096)

摘要:介绍了语音情感识别领域的最新进展和今后的发展方向,特别是介绍了结合实际应用的实用语音情感识别的研究状况。主要内容包括:对情感计算研究领域的历史进行了回顾,探讨了情感计算的实际应用;对语音情感识别的一般方法进行了总结,包括情感建模、情感数据库的建立、情感特征的提取,以及情感识别算法等;结合具体应用领域的需求,对实用语音情感识别方法进行了重点分析和探讨;分析了实用语音情感识别中面临的困难,针对烦躁等实用情感,总结了实用情感语音语料库的建立、特征分析和实用语音情感建模的方法等。最后,对实用语音情感识别研究的未来发展方向进行了展望,分析了今后可能面临的问题和解决的途径。

关键词:实用语音情感识别;情感计算;特征分析;情感模型;语料库;识别方法

中图分类号:TP391 **文献标识码**:A

Key Technologies in Practical Speech Emotion Recognition

Zhao Li, *Huang Chengwei*

(School of Information Science and Engineering, Southeast University, Nanjing, 210096, China)

Abstract: The latest and future progress in speech emotion recognition is introduced, especially in the practical speech emotion research considering the real world applications. The followings are mainly discussed: The history and development of affective computing research, the practical applications in affective computing, and the review of general speech emotion recognition methods, including emotion model, emotion database, feature extraction and emotion recognition algorithm. Considering the needs in real world applications, the key technologies in practical speech emotion research are focused on. Moreover, the current challenges in practical speech emotion recognition are analyzed, especially for the fidgetiness emotion, and the methods in database establishment, feature analysis and modeling techniques are reviewed. Finally, the outlook of future speech emotion research is given, and the future challenges and possible solutions are discussed.

Key words: practical speech emotion recognition; affective computing; feature analysis; emotion model; speech database; recognition method

引 言

信息技术正在越来越紧密地融入到人们的日常生活当中,人们需要便捷的获取信息,就需要同各类计算机进行交互。情感计算技术可以改进人们与高科技的交互方式,从传统的被动地使用机器,转变到自然地人机交互。情感是人类一种重要的本能,它同理性思维和逻辑推理能力一样,在我们的日常生活、工作、交流、处理事务和决策中扮演着重要的角色。随着计算机技术的发展,和谐的人机交互日益受到研究者的重视,它不仅要求计算机理解用户的情绪和意图,而且需要对不同用户、不同环境、不同任务给予不同的反馈和支持。人们试图创建一种能感知、识别和理解人的情感,并针对人的情感做出智能、灵敏、友好反应的计算系统,即

基金项目:国家自然科学基金(61273266;61231002;61375028;61301219;61301295)资助项目;教育部博士点专项基金(20110092130004)资助项目;江苏省自然科学基金(ZKJ201202)资助项目。

收稿日期:2013-12-27;**修订日期**:2014-01-20

赋予计算机像人一样地观察、理解和生成各种情感特征的能力，使计算机能够更加自动适应操作者。实现这些功能，首先必须要求能够识别操作者的情感，而后根据情感的判断来调整交互对话的方式。

美国 MIT 媒体实验室情感计算研究小组的领导人 Rosalind Picard 教授在 1997 年首次提出“情感计算”这个概念[1]，情感计算是一个高度综合化的技术领域，其研究内容包括：情感机理的理论研究、情感信号的采集、情感信号的分类、建模与识别、情感理解、情感表达及情感生成等几部分，主要从生理模式，面部表情及语音这 3 个切入点展开研究。目前，已有多个国家展开相关研究并取得了部分成果，研究机构不仅局限于各国科研院所，而且也得到了该国有关部门的关注。例如，美国的 MIT 媒体实验室的情感计算研究小组就在专门研究机器如何通过对外界信号的采样，如人体的生理信号（血压、脉搏、皮肤电阻等）、面部快照、语音信号来识别人的各种情感，并让机器对这些情感做出适当的反应。日本文部省将“情感信息处理的信息学、心理学研究”作为重点研究领域。我国中国科学院和国内众多高校在情感信息处理的研究上也取得了一定的进展[2]。

情感识别应用的一个著名的例子是 Rosalind Picard 教授提出来的“情感镜子”，情感镜子是一个与人交互的 Agent，可以帮助用户看到自己在不同的场合中的表现，如在准备面试或演讲中可以起到重要的作用。情感计算在人机交互中的应用具有广泛的应用前景。例如，在视频游戏领域，用户可以站在屏幕前进行虚拟的网球比赛。采用了情感识别技术后，用户的比赛体验可以获得极大的丰富。情感信息可以成为当前的多媒体内容识别与分析中的一个新的维度。电影或电视广播可以根据不同的情感内容来进行检索。在计算机辅助的教学中，情感计算技术可以帮助提高学生的学习兴趣。例如，当学生在学习过程中出现烦躁情绪时，通过情感识别技术，系统可以给予学生适当的鼓励或者是减慢学习进度。当学生感到枯燥乏味时，系统可以给出更具有挑战性的题目。在决策支持系统中情感识别技术同样能够起到重要的作用。例如，当操作人员表现出紧张或者烦躁等负面情绪时，系统可以给其分配较轻的任务。在人机交互中，引入语音情感技术后机器人或口语对话系统能够更加自然地与人进行对话[3]。在机器人研究领域，某些研究者正致力于开发具有类似人类能力的机器人，其中情感的理解与表达是一个关键的研究领域[4]。语音情感识别技术还可以应用于电话服务中心，系统可以检测谈话的语气和情感，从而提高服务质量。在信息查询系统中加入情感识别分析功能，可以根据用户情绪调整任务优先级，从而提高智能信息检索的效率；在教学实践环节，情感分析可以使得在教学的同时注重学生对信息的接收理解程度，从而及时调整教学节奏和进度，使得学生能更好地吸收知识；在工业生产领域，如电话通讯中，加入语音情感分析服务平台，可以进一步提高通信质量，使通话双方交流更通畅；在医学研究中，烦躁、焦虑、抑郁等不良情绪对治疗有很大的阻碍作用，如果能够更早发现病人情绪波动并及时稳定，对病人的康复也有着积极作用；在家居环境中，老年人面临“空巢”问题时，孤独和不被理解等问题都是影响他们安度晚年的重要因素，上班族没有足够的时间耐心与父母交流，如果能在家电系统中增加情感分析功能，使之在日常交互中能与老人形成一定情感交流，可以在一定程度上缓解老年人的精神空虚，上班族在一天的烦劳工作结束后，也可以从家庭环境中获得一定心理释放。

除了以上这些应用场合外，情感识别技术的研究本身能够对理解人类情感的产生、表达和感知具有重要的价值。随着计算机技术的日益进步，高技术越来越深入的融入到人们的日常生活中，自然、高效和人性化的交互技术已成为了一个极为重要的研究领域。

1 语音信号中的情感信息

语音作为人们交流的主要方式，语音信息在传递过程中由于说话人情感的介入而更加丰富。情感不仅可以强化语义信息，甚至可以改变语义信息。语音信号是语言的声音表现形式，情感是说话人所处环境和心理状态的反映，语音情感识别就是让计算机能够通过语音信号识别说话者的情感状态，是情感计算的重要组成部分[5-9]。由于语言包含了强烈的社会和文化背景，人们可以在非面对面的情况下表达出自己的心理状态，即使是不同肤色、不同语种的人在存在语言隔阂时，无法通过语义来沟通，但是仍然可以通过语音表达传递出情感信息并达到一些基本的理解。正是由于语言的这一社会文化特性，对语音信号中的情感进行分析判别时，不同语种之间所用的方法和判别标准既有共性又有异性，既具有一定参考性又不能完全照搬。这使得语音情感识别面临许多挑战性的难题，不仅存在于针对某种单一语言交流时的情感分析，也存在于不同文明背景下不同语种人们交流时的情感分析。

语音情感、面部表情、手势、姿态以及生理信号

等是情感表达与识别的途径。以上各种情感表达方式之间是如何相互影响的，至今还没有一个清楚的认识。Mehrabian 曾对情感和态度的表达中，非言语的表达方式的重要性做了研究。Mehrabian 的研究结果显示，在面对面的交互中，情感的表达具有 3 个基本的要素：语义、语气和身体语言。说话人给对方的好感程度与以上 3 个要素的关系可以粗略表示为：好感程度＝7％语义＋38％语气＋55％表情。

这一研究结果仅适用于语义与语气不相互冲突的情况，也就是说在说话人说反话讽刺时不适用。根据 Mehrabian 的这一研究结果，可以看到在语音交谈中，通常说话人的说话方式比说话内容具有更重要的情感交流的作用。虽然情感可以通过很多种途径表达，通过生理信号的测量来识别情感，如心电、脑电等，通常需要被试佩戴复杂的仪器设备，而语音作为情感的交流方式仍然是最便捷最自然的手段之一。语音作为日常生活中最常用的交流手段之一，特别是在同时处理多个事务的过程中，人机自然语音交互将发挥越来越重要的作用。

2　情感信息的定义和分类

在研究情感识别之前，需要做的第一件事就是定义所要研究的对象，从而明确研究的范围。然而“情感是什么?”这一个由来已久的问题，一直没有一个统一的答案。Scherer 曾指出：情感研究中的一个主要的问题是，缺乏对情感的一个一致的定义以及对不同情感类型的一个定性的划分。虽然在文学上对情感的描述，存在一些广泛接受的可能的分类，然而由于没有一个对情感描述的公认的方法，对情感的分类学研究也一直没有统一的意见[10-17]。

从日常表达上来说，人们常将情绪、情感、态度混淆起来，但是从研究的角度，研究者还是对此有不同看法。有研究者认为情感理解为一种因所处环境和心理状态而由主观冲动引起的强烈的感情状态，可以引起语音，表情以及行为上的表现。如 Klaus 对情感(Emotion)、情绪(Moods)、立场(Interpersonal stances)，态度(Atitude)和性情(Afect dispositions)这些近义词进行了研究，认为他们之间既有区别又不是绝对的独立，其中以情感和态度最容易混淆。Ohala 则认为态度更多是一种主动的感情色彩，而情感是被动的，两者之间有本质区别。然而到目前为止，研究者对情感的定义仍然没有达成一致的观点，Kleinginna 列举了近百名学者对情感的理解。对情感定义的不统一在一定程度上影响了情感计算的研究进展，这主要是由于情感随人类进化而不断发展的，人们对情感的认识在不同阶段不同切入点就呈现出不同的理解，不同的情绪机理学说应运而生。

早期的情绪研究主要是由哲学家、神经病学家、神经生理学家和心理学家分别进行的。早在公元前 5 世纪，古希腊学者就从生理心理角度，试图对情绪进行分析。赫拉克利特认为情绪状态是用身体温度、出汗量等一些生理参数来体现的，如对于一个正常状态下的人，他的身体温度偏冷，汗液的分泌偏干的。柏拉图将情感分作中性、高兴和痛苦 3 种状态，他认为高兴和痛苦两种状态是由中性状态分离而来；在中性状态时，人体的各个器官是和谐的，当这种和谐遭到破坏的时候，便产生了痛苦，而被破坏的和谐开始恢复时，便产生了高兴。亚里士多德则将高兴和痛苦看成是所有感情的基础，高兴是一种相对独立的情感，来源于中性的情感，而痛苦则是来源于高兴的反方向。如果没有感官和精神上的刺激，就不会有高兴和痛苦的存在。我国古代把情绪理解为人性的波动和扰乱，有“情，波也；心，流也；性，水也”《关尹子》；“性之有动者谓之情，性之有喜怒犹如水之有波浪”(程颐)，以及“性是未动，情是已动，心包括已，未动”(朱熹)之说。随着文明的发展，道德、宗教、生活等因素也被人们纳入情绪机理的考虑范围内，如斯多葛派(Stoic)禁欲主义者认为多数情绪是有害的，原因是人们有不正确的信念和不恰当的目标。佛教中对情也有论述，分别指的是“喜、怒、忧、惧、爱、憎、欲”七种情愫。中医中总结了“喜、怒、忧、思、悲、恐、惊”七种情绪状态，并指出这七种情态应该掌握适当。如果掌握不当，例如大喜大悲、过分惊恐等等，就会使阴阳失调、气血不周，从而这种精神上的错乱会演变到身体上，形成各种疾病。

总的来说，在情绪机理的研究发展中，比较有影响的情绪理论有以下几种：

(1) 詹姆斯-朗格情绪学说：美国心理学家詹姆斯和丹麦生理学家兰格分别提出内容相同的一种情绪理论。他们强调情绪的产生是植物性神经活动的产物。后人称它为情绪的外周理论。即情绪刺激引起身体的生理反应，而生理反应进一步导致情绪体验的产生。詹姆斯提出情绪是对身体变化的知觉。在他看来，是先有机体的生理变化，而后才有情绪。所以悲伤由哭泣引起，恐惧由战栗引起；兰格认为情绪是内脏活动的结果。他特别强调情绪与血管变化的关系。詹姆斯-兰格理论看到了情绪与机体变化的直接关系，强调了植物性神经系统在情绪产生中的作用；但是，他们片面强调植物性神经系统的作用，忽视了中枢神经系统的调节、

控制作用，因而引起了很多的争议。

(2) 丘脑情绪学说：又称为坎农-巴德学说，它反驳了詹姆斯-朗格情绪学说，丘脑情绪学说认为情绪的产生是大脑皮层解除丘脑抑制的综合功能，即激发情绪的刺激由丘脑进行加工，同时把信息输送到大脑及机体的其他部分。输送到大脑皮层的信息产生情绪体验；输送到内脏和骨骼肌的信息激活生理反应。身体变化和情绪经验是同时发生的，而情绪感觉则是由大脑皮层和自主神经系统共同激起的结果。情绪发生的中心不是外周神经系统，而是丘脑。此后的一些实验也证明，情绪的复杂生理机制在很大程度上取决于下丘脑、边缘系统、脑干网状结构的功能，大脑皮层调节情绪的进行，控制皮层下中枢的活动。

(3) 认知-评价学说：Arnold 等人认为情绪是驱利避害的一种体验倾向，任何评价都带有情绪的性质，评价是由知觉而产生的活动倾向，当倾向强烈时就可称为情绪。对情境事件的评价而引起的情绪会诱导人选择适合于情境的反应行动。该学说又被扩展为评价、再评价过程，包括筛选信息、评价、以及应付冲动、交替活动、身体反应的反馈、对活动后果的知觉等成分。他认为情绪是一种综合性的行为反应，每种情绪都包括生理、行为和认知 3 种成分反应。这 3 种成分在每种特定的情绪中各自起着不同的作用，相互作用、互为因果。它们的不同组合是构成各种具体情绪模式的特定标志。

(4) 动因-分化学说：Tomkins 等人认为情绪是以身体为基础，对某些动因体系(Motivational system)所做的放大。动因-分化学说比认知学说更注重情绪的作用，情绪是认知发展的契机，人完全可以由各种情绪激动起来，以激起人去认知和行动。

(5) 认知-生理学说：是詹姆斯-朗格情绪学说和认知学说的结合，认为个人对自己情绪状态的认知性解释是构成情绪的主要因素，经刺激所激活的生理变化是构成情绪的次要因素，泛化的生理反应决定情绪经验的强度，而情绪的性质则由对情境的知觉所决定。

对情感定义的不唯一性，使得在情感的分类问题上也存在分歧。前期研究者认为对语音情感的研究就是找出一个基本的情感类型列表，然后再研究表中的情感是如何在人类语言交流时表现出来。由此发展出两种情感分类观点：基本情绪论(Basic emotion theory)和调色板情绪论(Palette theory of emtion)。前者认为存在一些情绪状态是基本的纯粹的，剩余情感则是次要的不单纯的，这种观点比较符合现代心理学认知，它将情感看成是由分立的基本情感组成，每种类型各有其独特的体验特性、生理唤醒模式和外显模式；后者认为除去那些基本纯粹的情绪状态外，其他情感是在单纯情感的基础上衍变而来，就像调色板调色一样。不同研究者提出的情感类型从 2 种到近百种不等，中国古代就对情感分成了 7 类，就是常说的七情六欲中的七情，在《礼记·礼运》中解释为："喜、怒、哀、惧、爱、恶、欲七者弗学而能"。而中医没有把"欲"列在七情之中，换为了"喜、怒、忧、思、悲、恐、惊"。西方一些研究者的情感类型情感类型列表如表 1 所示。

表 1 基本情感分类列表

Table 1 List of the basic emotion categories

研究者	基本情绪定义
Plutchik	接纳，愤怒，期待，厌恶，喜悦，恐惧，悲伤，惊奇
Arnold	愤怒，厌恶，勇气，心情低落，欲望，绝望，恐惧，仇恨，希望，爱情，悲伤
Ekman/Friesen/Ellsworth	愤怒，厌恶，恐惧，快乐，悲伤，惊奇
Frijda	欲望，快乐，兴趣，惊奇，惊奇，悲伤
Gray	愤怒，恐惧，焦虑，喜悦
Izard	愤怒，轻蔑，厌恶，痛苦，恐惧，内疚，兴趣，快乐，羞愧，惊喜
James	恐惧，悲伤，爱情，愤怒
McDougall	愤怒，厌恶，得意，恐惧，屈从，柔情，难怪
Mowrer	痛苦，快乐
Oatley/Johnson-Laird	愤怒，厌恶，焦虑，悲伤，幸福

近 20 年，在坐标空间中对情感定位成为另一个情感分类研究热点，称之为维度空间论，主要是集中在二维论和三维论中。二维论是指效价维/快乐维(Valence/hedonic tone)和激活维/唤醒维(Activation/arousal)；三维论主要是在二维论的基础上增加一个控制维/姿态维(Control/stance)。其中效价维主要体现为情感主体的情绪感受，表示情感的积极或消极程度，喜欢或不喜欢程度，正面或负面程度，话者借助情感要表达的就是他对人或事物的喜欢程度和积极或消极的态度；激活维是指与情感状态相联系的机体能量激活的程度，是对情绪的内在能量的一种度量，表征个体对于各种活动的参与性，是活跃的还是呆板的，是兴奋的还是冷淡的；控制维体现的是主体对情感状态的主观控制程度，用以区分情感状态是由主体主观发出的还是受客观环境影响产生的，比如轻蔑和恐惧，就处于控制维度不同的两端。

Russel 等人通过激活效价空间上用一个情感轮(Emotion wheel)对情感进行分类[18]，图 1 所示的是情绪的二维模型。情感分布在一个圆形的结

构上，结构的自然原点认为是一种具有各种情感因素的状态，但是由于这些情感因素在该点的强度太弱而得不到体现。通过向周围不同方向扩展，表现为不同情感。情感点同原点的距离体现了情感强度。相似的情感相互靠近，相反的情感则在二维空间中相距 180 度。在这个二维空间中加入了强度做为第三个维度后，可以得到一个三维的情感空间模型。如图 2 所示。以强度、相似性和两极性划分情绪，模型上方的圆形结构划分为 8 种基本情绪：狂喜、警惕、悲痛、惊奇、狂怒、恐惧、接受和憎恨，越邻近的情绪性质上越相似，距离越远，差异越大，互为对顶角的两个扇形中的情绪则是相互对立的。圆形结构的中心为自然原点。在强度上延伸为三维椎体，强度越弱，情绪的兴奋度越低，越消极，反之则兴奋度越高越积极。

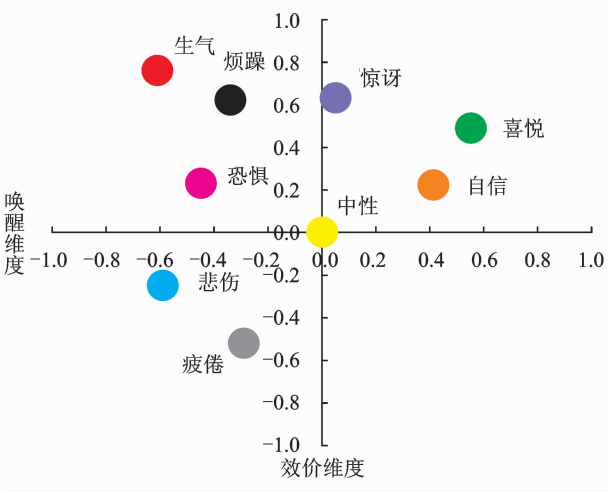

图 1　情绪的二维模型

Fig. 1　Two-dimensional emotion model

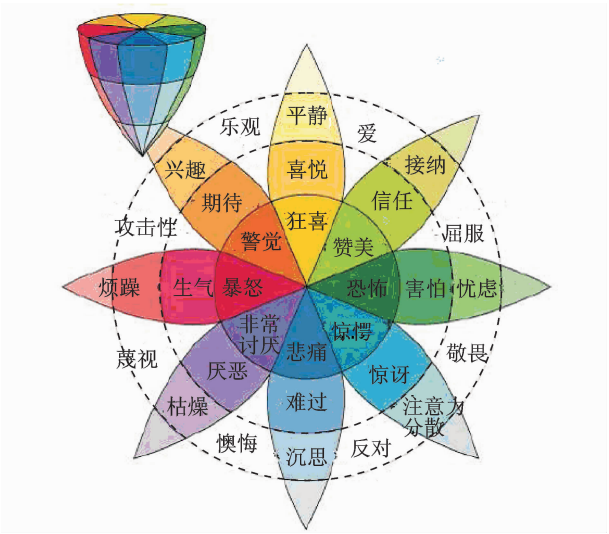

图 2　Plutchik 三维情绪模型

Fig. 2　Plutchik's three-dimensional emotion model

3　语音情感信息数据库的建立方法

情感语音数据库是进行语音情感分析的前提条件。根据数据获取途径，目前国际上语音情感研究人员所用的数据按照获取途径大致可分为 4 类：表演数据、激励数据、启发数据和摘引数据。按照语料自然度，数据可分为模仿数据、诱发数据和自然数据[19-30]。

表演数据主要是说话人用表演方式朗读某条有情感要求的给定语句，同时进行录音获得数据。一般要求说话人是受过专业训练的演员、播音员。这类数据的优点是，在录制的时候可以根据研究需求随时调整数据的录制，满足性别、文字和情感等要求；而且录音人员大多为受过表演训练的演员，所录语音具有明显的情感表现力，在其后的数据有效性交叉测听验证时，具有较高的可识别性。同时，此类数据的缺点也是很明显的，由于是由专业人士表演获得的，数据的情感表现具有一定的夸张度，不同演员对度的把握也不尽相同，人们在日常交流中情感的正常流露与表演出的情感还是有一定距离的，当把根据表演数据得到的情感分析方法在用于日常交流的语音时，会出现一定偏差，不利于日后研究。

由于表演数据的真实度不高，情感的表达不受心理活动刺激，研究人员在进行录制前，先设定一个有情感倾向的场景文本让说话人朗读，用文本的内容来激励说话人情感，通常情况下场景文本较长，说话人在朗读的过程中，心理发上变化而使语音逐渐带有情感。这种数据就称之为激励数据。有时场景文本也可由图片影像等其他方式激励说话人。其优点是符合人类情绪产生的过程，能够体现出情感的渐变性，真实度较高。其缺点是场景文本内容的情感倾向可能会影响说话人对语音的判断，而这些影响无法通过分析语音特征剔除。

启发数据是通过多人之间的交流获得的，通常是在一个自由的环境下，由一个或多个启发者人员与被录音人员进行交流，交流时间较长，谈话过程中启发人员通过对被录音者的了解，随时调整话题和控制交流速度，启发出后者的情感。启发过程中也可借助其他非语言类工具启发被录音人情感。较之前两类数据，启发数据真实度较高，由于录音时间较长，按照人类情绪发生过程，数据前期较为平稳，进入中期，话者逐步进入某种情绪状态并最终到达情绪高潮，后期又逐渐归于平静状态。但是此类数据也是对启发人员要求较高，不仅要对被录音人有一

定程度的了解，能够把握说话人的心理变化，而且需要很强的现场调控能力。启发数据的另一个缺点就是由于被录音人的个体差异性，实际录音过程中，可能不会出现一些极端的情感，如暴怒、狂喜等。

摘引数据主要是指从广播电视日常生活中截取我们感兴趣的包含情感的语音片断。这种方法的优点是情感的真实度较之表演数据更高，情感表达直接由心理状态触发而成，有上下文内容关联信息；数据来源丰富，并且截取的多媒体素材中不仅仅是音频信息，这对日后发展多模态情感识别研究提供帮助。但是由于数据的海量，寻找合适的数据需要花费大量的时间和人力，对存在背景音的语音材料还需进行分离预处理等步骤才能得到干净的情感语句。另外，对某些数据来源可能还存在版权等法律问题，这些都是摘引数据的劣势所在。

模仿数据是用专业的、蓄意的方式产生情绪的语音，一般由专业演员表演产生。而诱发数据是由于劝诱产生情绪的语音。诱发数据的自然度介于模仿数据和自然语音数据之间，一般是由非专业的普通人在文字，图片等方式的诱导下获得的。自然语音数据则直接取材于人们日常交流中。

总的来说表演数据和部分激励数据都是通过表演者模仿获得的。真实性不高，但是在实验室环境下，易于研究，有较高识别率。也是用的最多的建库途径。启发数据的真实度较高，对启发者要求较高，存在一定不可预知性，但是仍然获得了部分研究者的认可。启发数据和部分激励数据属于诱发获得。摘引数据既有从影视剧中获得的，也有从访谈日常谈话中获得的，因此，数据的自然度涵盖范围最大，但较之单纯的表演数据，由于有上下文的关联，情感表达上还是有一定优势。

不论是通过哪种途径获得情感语音数据，建立一个完备的语音数据库都是可持续研究的关键所在。完备性要求是指，要符合语言的概率模型，在保证文本真实性和话语自然度的前提下，用尽可能少的语句来覆盖所有的汉语发音现象，即包含所有合理的音联关系，包含各种音节内和音节间的元辅音搭配关系，能体现协同发音现象及发音的韵律特征，能体现汉语语音学、声学的各种特征。情感语音库的完备性要求比较特殊，与其他的语音库的完备性要求不尽相同。情感语音库采集的是情感，要求每种情绪类别的语音数据，包含该情绪的各种可能的情感表达方式。由于情感的表达受主观因素影响较大，不同的谈话人有不同的情感表达习惯。因此，保证说话人的数量达到一定的规模，选择表演、激励等各种不同的采集方式，设计充足的语句文本等，都有助于建立完备的情感语音库。

4 语音情感信息特征的分析与提取

情感语音当中可以提取多种声学特征，用以反映说话人的情感行为的特点。情感特征的优劣对情感最终识别效果的好坏有非常重要的影响，如何提取和选择能有效反映情感变化的语音特征，是目前语音情感识别领域最重要的问题之一。在过去的几十年里，针对语音信号中的何种特征能有效的体现情感，研究者从心理学、语音语言学等角度出发，作了大量的研究。许多常见的语音参数都可以用来研究，这些语音参数也常用于自动语音识别和说话人识别当中。例如：短时能量、过零率、有声段和无声段之比、发音持续时间、语速、基音频率、共振峰频率和带宽、镁尔倒谱参数(Mel-frequency cepstral coefficients，MFCC)等等。

近年来，在情感特征的分析过程中，研究者们开始关注到语料的真实度问题，以前的表演语料具有一定的夸张成分，在此类语料上获得的情感特征与实际情况可能存在一定的偏差。

在过去的情感特征分析中，存在的最大的问题是不同研究者之间的实验结果具有较大的差别，由于语料库的不统一，研究成果之间的可比性较差。往往在一个数据库上行之有效的特征，迁移到另一组语料上就不能获得同样的性能。因此，在今后的研究中，应该关注跨数据库的扩展性能的研究，对不同民族之间和不同语种之间的情感表达的差异应该受到研究者的重视。

此外，对于特殊人群和特殊工作环境中的情感特征的分析，具有较高的实际意义，应当受到重视。例如，高压环境下人员的情感和心理状态变化，狭小密闭环境引发的负面情绪，这些都是值得研究的课题。可以预期，在实际环境中引发的情感状态，其特征应该与标准数据库当中的基本情感类别的特征有所不同。因此对实用语音情感特征的研究具有较高的实际意义。

4.1 情感特征的构造

用于识别和建模的特征向量一般有两种构造方法，静态统计特征和短时动态特征。动态特征对音位信息的依赖性较强，为了建立与文本无关的情感识别系统，本文中选用了静态统计特征，如表 2，3 所示。

表 2　情感语音的基本声学特征构造(上)

Table 2　Construction of the basic acoustic features of speech emotion (A)

特征编号	特征名称
1～6	短时能量的均值、最大值、最小值、中值、范围和方差
7～12	短时能量一阶差分的均值、最大值、最小值、中值、范围和方差
13～18	短时能量二阶差分的均值、最大值、最小值、中值、范围和方差
19～24	基音频率的均值、最大值、最小值、中值、范围和方差
25～30	基音频率一阶差分的均值、最大值、最小值、中值、范围和方差
31～36	基音频率二阶差分的均值、最大值、最小值、中值、范围和方差
37～42	过零率的均值、最大值、最小值、中值、范围和方差
43～48	过零率一阶差分的均值、最大值、最小值、中值、范围和方差
49～54	过零率二阶差分的均值、最大值、最小值、中值、范围和方差
55	语速
56～57	基音频率一阶抖动、基音频率二阶抖动
58～61	0～250 Hz 频段能量占总能量的百分比、0～650 Hz 频段能量占总能量的百分比、4 kHz 以上能量占总能量的百分比、短时能量抖动
62～65	发音帧数、不发音帧数、不发音帧数和发音帧数比、发音帧数和总帧数比
66～69	发音区域数、不发音区域数、发音区域数和不发音区域数之比、发音区域数和总区域数之比
70～71	最长发音时间、最长不发音时间

表 3　情感语音的基本声学特征构造(下)

Table 3　Construction of the basic acoustic features of speech emotion (B)

特征编号	特征名称
72～77	谐波噪声比(HNR)的均值、最大值、最小值、中值、范围和方差
78～83	0～400 Hz 频段内谐波噪声比的均值、最大值、最小值、中值、范围和方差
84～89	400～2 000 Hz 频段内谐波噪声比的均值、最大值、最小值、中值、范围和方差
90～95	2 000～5 000 Hz 频段内谐波噪声比的均值、最大值、最小值、中值、范围和方差
96～101	第一共振峰频率(F_1)的均值、最大值、最小值、中值、范围和方差
102～107	第二共振峰频率(F_2)的均值、最大值、最小值、中值、范围和方差
108～113	第三共振峰频率(F_3)的均值、最大值、最小值、中值、范围和方差
114～119	第四共振峰频率(F_4)的均值、最大值、最小值、中值、范围和方差
120～125	第一共振峰频率一阶差分的均值、最大值、最小值、中值、范围和方差
126～131	第二共振峰频率一阶差分的均值、最大值、最小值、中值、范围和方差
132～137	第三共振峰频率一阶差分的均值、最大值、最小值、中值、范围和方差
138～143	第四共振峰频率一阶差分的均值、最大值、最小值、中值、范围和方差
144～149	第一共振峰频率二阶差分的均值、最大值、最小值、中值、范围和方差
150～155	第二共振峰频率二阶差分的均值、最大值、最小值、中值、范围和方差
156～161	第三共振峰频率二阶差分的均值、最大值、最小值、中值、范围和方差
162～167	第四共振峰频率二阶差分的均值、最大值、最小值、中值、范围和方差
168～171	第一到第四共振峰频率的一阶抖动
172～175	第一到第四共振峰频率的二阶抖动
176～181	第一共振峰带宽的均值、最大值、最小值、中值、范围和方差
182～187	第二共振峰带宽的均值、最大值、最小值、中值、范围和方差
188～193	第三共振峰带宽的均值、最大值、最小值、中值、范围和方差
194～199	第四共振峰带宽的均值、最大值、最小值、中值、范围和方差
200～205	第一共振峰带宽一阶差分的均值、最大值、最小值、中值、范围和方差
206～211	第二共振峰带宽一阶差分的均值、最大值、最小值、中值、范围和方差
212～217	第三共振峰带宽一阶差分的均值、最大值、最小值、中值、范围和方差
218～223	第四共振峰带宽一阶差分的均值、最大值、最小值、中值、范围和方差
224～229	第一共振峰带宽二阶差分的均值、最大值、最小值、中值、范围和方差
230～235	第二共振峰带宽二阶差分的均值、最大值、最小值、中值、范围和方差
236～241	第三共振峰带宽二阶差分的均值、最大值、最小值、中值、范围和方差
242～247	第四共振峰带宽二阶差分的均值、最大值、最小值、中值、范围和方差
248～325	0～12 阶镁尔倒谱参数(MFCC0-MFCC12)的均值、最大值、最小值、中值、范围和方差
326～403	0～12 阶镁尔倒谱参数一阶差分的均值、最大值、最小值、中值、范围和方差
404～481	0～12 阶镁尔倒谱参数二阶差分的均值、最大值、最小值、中值、范围和方差

文本的变化会对情感特征有较大的影响。情感语音当中大致包含 3 种信息来源:说话人信息、语义信息和情感信息。在构造情感特征和选择特征时,不仅需要使得特征尽可能多地反映出情感信息,也就是随着情感的变化而发生明显的变化,而且还需要尽量保持特征不受到语义变化的影响。

4.2 特征的降维方法

上文提取了大量的基本声学特征,由于受到训练样本规模的限制,特征空间维度不能过高。特征降维,在一个模式识别系统中具有重要的作用。原始的基本特征或多或少地能够提供可利用的信息,来增加类别之间的可区分度。从信息的增加的角度来说,原始特征的数量应该是越多越好,似乎不存在一个上限。然而,在具体的算法训练当中,几乎所有的算法都会受到计算能力的限制,特征数量的增加,最终会导致"维度灾难"的问题。以高斯混合模型为例,它的概率模型的成功训练依赖于训练样本数量、高斯模型混合度、特征空间维数三者之间的平衡。如果训练样本不足,而特征空间维数过高的话,高斯混合模型的参数就不能准确地获得。

本文对上文中列出的所有基本声学特征进行特征降维,既能够反映出这些特征在区分情感类别上的能力,又是后续的识别算法研究的需要。总结语音情感识别领域近年来的一些文献,研究者们主要采用了以下的一些特征降维的方法:线性鉴别分析(Linear discriminant analysis,LDA)、主成分分析(Principal components analysis,PCA)、Fisher 鉴别率(Fisher discriminant ratio,FDR)、序列前向选择(Sequential forward selection,SFS)等。其中,SFS 是一种封装器方法(Wrapper),它对具体的识别算法依赖程度比较高,当使用不同的识别算法时,可能会得到差异很大的结果。

在特征维数较高时,LDA 的压缩性能是非常明显的。然而在实际中 LDA 的应用会受到训练数据量的限制,当原始特征维数非常高,而训练数据量不足时,会导致矩阵出现奇异值,LDA 无法正常使用。因此,在本文中处理高维数据时,可以采用 PCA 进行第一步降维,然后再使用 LDA 降维。

5 语音情感特征的统计模型与识别算法

语音情感识别在人机自然交互领域中有着重要的应用前景。在不久的将来,智能家用电器、智能服务型机器人等智能工具要进入到人们的日常生活中,必然要面临的问题是人与机器如何交互的问题。在以个人为中心的服务中,包括个人电脑等消费电子,普通大众并不习惯于键盘、鼠标等生硬的操作方式,而语音是人类最自然、最便捷的交流方式之一,以语音、表情、手势等自然的方式与机器沟通已成为了人机交互研究领域的一个趋势。使智能机器具有理解人类情感的能力,识别用户的喜悦、烦躁、满意、愤怒、急切等情感,具有重要的实际意义。通过语音情感识别,在人机语音通信中获取情感等非语义信息,能够使得智能机器具备"察言观色"的能力,能够适应各种实际的社会场合,准确地理解用户的意图,自然地与用户进行沟通。

语音情感识别是以情感机理研究为基础、在获取了有效的情感语音信号后,将情感信号与情感机理相应方面的内容对应起来,对所获得的信号进行建模和识别。情感机理研究主要指对情感状态判定及其与生理和行为之间的关系。涉及到心理学、生理学、认知科学等方面学科。情感信号的获取研究主要是指各类有效传感器的研制,它是情感计算中极为重要的环节,这里主要是各类高性能的录音仪器。通过对录得的语音信号进行交叉验证其有效性后,对信号进行建模和识别。例如,隐马尔可夫模型(Hidden Markov models,HMM)、贝叶斯等模型就被广泛采用并加以改进,取得了一定的识别效果[31-41]。

这里简要总结了各种现有的语音情感信息的统计模型与识别算法,如表 4 所示。模式识别领域中的诸多算法都曾用于语音情感识别的研究,典型的有 HMM、高斯混合模型(Gaussian mixture model,GMM)、支持向量机(Support vector machine,SVM)和人工神经网络(Artificial neural network,ANN)等,表 4 中初步比较了它们各自的优缺点以及在部分数据库上的识别性能。

GMM 是一种拟合能力很强的统计建模工具。GMM 的主要优势在于对数据的建模能力强,理论上来说,它可以拟合任何一种概率分布函数。而 GMM 的主要缺点,也正是对数据的依赖性过高。因此在采用 GMM 建立的语音情感识别系统中,训练数据的特性会对系统性能产生很大的影响。

GMM 在说话人识别和语种识别中获得了成功的应用。就目前来说,很多研究的结果显示,GMM 在语音情感识别中是一种较合适的建模算法。近年来的研究文献中,报道了不少采用 GMM 建立的语音情感识别系统。这些基于GMM的识

表 4　各种识别算法在语音情感识别应用中的特性比较

Table 4　Comparison of the characters of various recognition algorithms in speech emotion recognition

识别算法	对语音情感数据的拟合性能	识别率	优点	缺点
GMM	高	在 AIBO 数据库、本文数据库上表现较高	对数据的拟合能力较高	对训练数据依赖性强
SVM	较高	在柏林库上表现较高	适合于小样本训练集	多类分类问题中存在不足
KNN	较高	在柏林库上表现一般	易于实现,较符合语音情感数据的分布特性	计算量较大
HMM	一般	在柏林库上表现较高	适合于时序序列的识别	受到音位信息的影响较大
决策树	一般	在 AIBO 数据库上表现一般	易于实现,适合于离散情感类别的识别	识别率有待提高
ANN	较高	在日语情感语音上表现一般	逼近复杂的非线性关系	容易陷入局部极小特性和算法收敛速度较低的
混合蛙跳算法	较高	在汉语音情感数据上表现较高	优化能力强,有利于发现情感数据中潜在的模式	在迭代后期容易陷入局部最优,收敛速度较慢

别系统,相对于其他识别算法来说获得了较好的识别率。在 2009 年,语音领域的著名的国际会议(Interspeech)上,举行语音情感识别的评比。基于 GMM 的识别系统在总体性能上获得了该次比赛的第一。

采用何种建模算法最适合语音情感识别,一直是研究者们非常关注的问题。本文认为,在不同的情感数据库上、不同的测试环境中,不同的识别算法各有优劣,对此不能一概而论。然而,目前研究者们对自然语料非常重视,在自然语料中的情感模式较为复杂,不同的说话人、不同的性格特点、不同的上下文环境等等因素都会增加数据的复杂度。高斯混合模型对这些数据的适应能力较强,可能是多数应用场合的一种合理选择。

在模式识别方面,各国研究人员在语音情感信息处理领域几乎利用了所有的模式识别手段,新方法的应用和对比层出不穷。模式识别方法大致可分为 3 大类:模板匹配法、概率统计法、辨别分类器法。其中模板匹配法代表性的有动态时间规整法(Dynamic time warping,DTW)和矢量量化方法两种;概率统计法代表性的有 HMM 方法和 GMM 方法两种;辨别分类器法如 ANN 方法和 SVM 方法。此外,把以上方法与不同特征进行有机组合,即混合方法也是情感识别中常见的,如 GMM/SVM 混合模型方法、SVM/HMM 混合模型方法等等[36-40]。

Yamada 等[42]对将神经网络应用于提取语音中的情感进行了研究,这些情感包括悲伤、兴奋、欢乐和愤怒。对于这些基本的人类情感,运用神经网络可以达到 70% 的识别率。Nicholson 所研究的系统的整个神经网络由 8 个子网构成,每个子网处理一种特定的情感。测试发现,负面的情感,比如愤怒和悲伤容易识别,但正面的情感,比如喜悦,不易识别。Tato[33]等人使用 SVM 作为分类器对四类(喜、怒、悲、平常)情感进行识别研究,最后实现了 73% 的平均识别率。Tin Lay Nwe 等[43]采用了 Mel 频率(Mel-frequency)语音能量系数和 HMM 分类方法,这种方法能够比较有效地识别出语音所包含的情感,但还不足以反映情感的细节,对情感进行精确的区分。赵力等[44-47]分别采用 PCA,HMM,GMM,QDF 等方法进行识别,也取得了 70%～90% 的识别率。

6　语音情感识别系统中的难点

目前基于语音的情感识别系统中还存在不少困难,离实际应用的要求还有一定的距离。用于识别语音情感的机器学习算法通常需要大量的训练数据。在相对成熟的语种识别或说话人识别领域内,训练一个正常工作的系统通常需要几百个小时的语音数据。标注后的语音情感训练数据是稀疏的,这对情感识别研究带来了难题。目前的情感识别研究缺乏足够的标注好的自然情感语音数据。另一方面,表演语音相对容易获取,但是用表演语音数据代替自然语音数据会带来系统性能的下降,研究表明表演情感数据与真实的情感数据之间有着较大的差异。然而现实世界中的情感在某种程度上也是在各种因素的影响下表达的,在不同的社会环境下都会带有一定程度的掩饰和表演的成分。

采集充分多的自然情感语音数据具有一定的困难，大部分的真实情感出现在特定的社交场合，在自然对话中出现的情感会受到观察者的影响，在实验室里很难进行完全真实的重现。当人们获知他们的对话在被采集和录制时，情感的表达会受到一定程度的抑制。例如，在 Ekman 的研究中，日本人会在参与实验中用微笑来掩盖负面的情绪。对语音情感来说，说话本身是一个受到高度控制和约束的过程，不少受控较少的情感表达需要一些极端的事件来激发，在进行情感语音的采集过程中伦理道德也是不可忽视的因素，被试往往出于隐私的考虑而不会给出最真实的情感表达。

当获得了自然的语音情感数据后，下一步就需要来描述语音中出现的这些情感。对自然情感的标注是一件困难的工作，特别是在上下文场景未知的情况下要准确地判断出说话人的情感更加困难。而且对情感的表达和感知某种程度上是因人而异的，不同的人对情感的表达能力不同，对同一段情感语料也存在不同的感受。因此不得不采用大多数人投票的方案来进行情感的标注，当多人对一段语料有相同的标注时将其作为基准。对于同一个标注人，还需要考虑其给出的判断的可靠程度。对语音情感数据的预处理是一个需要大量人力和时间的过程。

以往的研究表明，声学特征对区分不同的情感类别有重要的作用。激活维上差异较大的情感，如愤怒(高激活度)和悲伤(低激活度)，通过声学特征能够得到较好的区分。然而在愉悦维度上对喜悦和愤怒的区分则较为困难。虽然近年来大量的音质特征被用来区分正面和负面的情感，然而离实际应用的要求还有一定的距离。因此，在语音情感特征分析中，尽可能多地提取声学特征，用特征选择算法来选取区分性最高的特征。与自动语音识别和人脸表情识别领域不同，寻找一套有效的声学特征以及配套的识别算法的研究还没有得到一个统一的结论，目前广泛使用的语音情感特征和识别算法，还不能很好地捕获自然语音中的不明显的情感表达。而对于表演语音的情感区分性能较高，是由于表演语音情感较为强烈，在激活度上的差异较大。

虽然世界各国的研究人员在语音情感研究的领域取得了许多研究成果，但是整个语音情感信息处理领域还处在一个较低的水平。特征提取的手段极其局限，对于模式识别的手段，虽然有不同的应用方法，但是由于研究项目中使用的数据各异，而使得这些文献间类比的可能性不大。纵观近几年语音情感文献的研究结果，不仅它们的语音数据库不同，而且不同识别算法的应用也造成了高低不等的识别率。

在语音情感信息处理领域，无论是特征的提取，还是模式的识别，都存在相当多的问题。总结起来有如下几类：

(1) 目前国内外对情感识别的研究，主要集中在几类基本情感的识别上，尚不能满足实际应用中的需求，缺乏实用语音情感的数据库以及在此基础上的特征分析与识别的研究。

(2) 没有一个统一的共享的情感数据库用于语音情感识别，由于研究项目中使用的数据各异，而使得各类研究文献间类比的可能性不大。而且由于语种的关系，不同语种之间的研究成果的交流也存在一定障碍。

(3) 在情感特征参数的提取和选择上，特征提取的手段极其局限，几乎所有的研究人员都是采用韵律特征或者这些韵律特征的线性组合和变换作为研究对象。虽然少数研究人员也提出了一些新的特征参数，但是所有这些成果目前还停留在研究阶段，对其的广泛认可仍需时间。

(4) 情感识别算法的使用上，纵观近几年语音情感文献的研究结果，由于语音数据库不同，使得不同识别算法的应用造成了高低不等的识别率。对某些算法的有效性上仍存在验证问题。

7 实用语音情感识别的研究现状

语音情感识别是实现以人为中心的自然人机交互的关键技术之一，近年来受到了来自计算机科学、心理学、认知科学与行为科学等各个领域的研究者们越来越高的关注。情感状态的识别与在此基础上的心理评估具有很高的实际应用价值，特别是在载人航天等军事领域中，长时间的、单调的、高强度的任务，会使得相关人员面临严酷的生理以及心理考验，引发某些负面的情绪。探讨这些情绪对工作能力的作用及其机制和影响因素，具有非常重要的应用价值，可以研究提高个体认知和工作效率的方法、避免影响认知和工作能力的因素。然而以往的语音情感识别，集中在对几种基本情感的研究上，实验手段上往往采取表演的方式来模仿实际环境中的真实情感。通过对基本的几类语音情感的分类研究，虽然能够在理论上验证各种识别算法的优劣性能，能够用于寻找对识别基本情感类别有效

的声学特征,但是仅停留在对基本情感类别的研究上,远远不能满足实际应用中的要求。

人员的心理素质(如情绪稳定等)是实际任务中仪器和装备所无法替代的关键因素,直接关系到航空航天等任务的顺利完成。要保持良好的情绪状态,除了进行专业的心理训练、任务执行过程中的心理干预以外,配套的情绪检测仪器的研制是必要的硬件基础,是对情绪评价提供客观指标的依据。因此实时地在线情绪状态评估,以及在此基础上的情绪能力的考核,具有非常重要的实用意义。然而目前国内外对情感识别的研究,主要集中在几类基本情感的识别上,尚不能满足实际应用中的需求。由于实际应用中的需求,语音通话中"烦躁"情感具有重要的研究价值。因此,对烦躁情感的识别是语音情感识别中非常重要的一项研究内容,具有重大的实际意义。

在实际的语音情感识别应用中,还面临着情感语料真实度的问题。根据 Scherer 的观点,人类声音中蕴含的情感信息,受到无意识的心理状态变化的影响,以及社会文化导致的有意识的说话习惯的控制。语音情感中的这种无意识和有意识控制对情感识别在实际中的应用至关重要。然而在目前的语音情感数据的采集中,广泛使用的是表演的方式,在实际的语音通话和自然交谈中,说话人的情感对语音产生的影响,常常不受说话人控制,通常也不服务于有意识的交流目的,而是反映了说话人潜在的心理状态的变化。相反,演员能通过刻意地控制声音的变化来表演所需要的情感,这样采集的情感数据对于情感语音的合成研究没有问题,但是对自然情感语音的识别研究不合适,因为表演数据不能提供一个准确的情感模型。为了能更好地研究实际环境中的情感语音,有必要采集除表演语音以外的、较高自然度的情感数据。根据自然程度和采集方法,情感语料可以分为自然语音、诱发语音和表演语音 3 类。表演语料的优点是容易采集,缺点是情感表现夸张,与实际的自然语音有一定的差别,因此导致表演数据的可靠性较差。基于表演情感语料建立情感识别系统,会带入一些先天的缺陷,这是由于用于识别模型训练的数据与实际的数据有一定的差别,导致了提取的情感特征上的差别。因此,以往基于表演语料的识别系统,它的情感模型在实验室条件下符合样本数据,在实验测试中也能获得较高的识别率;但是在实际条件下,系统的情感模型与真实的情感数据不能符合得很好,导致了识别正确率的显著下降。因此需要通过心理学实验的方法来采集实用语音情感的诱发数据,尽可能地使训练数据接近真实的情感数据。

实用语音情感数据库的建立,是实用语音情感的研究基础,具有极为重要的意义。目前国际上流行的语音情感数据库有丹麦语数据库、柏林数据库、Groningen ELRA 数据库、Reading-leeds 数据库、ESP 数据库和 Amir 数据库等,中文语音情感数据库有中国科学院自动化所的 CASIA 语料库、中国社会科学院录制的 CASS-ESC 等数据库。然而现有的这些语音情感数据库主要通过表演的方式采集几类基本情感类别的语音数据,不能满足实用语音情感研究需要。在语音情感识别的实际应用中,对建立情感模型所用的情感数据的真实性要求特别高,以往基于表演数据训练得到的模型,虽然在实验室条件下能够通过识别测试,但是在实际环境中对真实情感数据的识别性能较差。面向实际应用的这一特点,决定了实用语音情感数据库必须要保证语料的真实性,而不能采用传统的表演方式采集数据。针对这一问题,通过实验心理学的手段,在计算机游戏创造出的虚拟的情景中诱发被试说出带有特定情感的话语,能够采集较高自然度的情感数据。

在实用语音情感的特征分析中关注最多的是韵律特征和音质特征。心理学和语言心理学的研究人员提供了大量的关于语音学和韵律学的研究结果,可以用来提取特征。一般情况下,语音的情感相关性的表示形式可以通过说话人模型或者声学模型来实现。有研究者认为语音情感识别的重点在韵律特征;而随着研究的深入,另外一些研究者认为,语音特征和韵律特征相互结合才能表达情感,仅有韵律特征是不可能表达情感的。到目前为止,已有的研究成果表明,针对情感识别所采用的特征大多是韵律特征,也就是超音段特征,如基音、强度、持续时间、以及它们的衍生参数,主要是统计参数,如均值、方差、中值、最大最小值、轮廓变化等。语音音质听觉方面的信息也是常常需要考虑的因素。一些特定元音在结构上的变化直接依赖于情感,而另一些元音则依赖于句子中的位置及话者是否用错了重读模式。音质类特征中代表性的有:共振峰,MFCC,LPCC,PLP 等。韵律特征和音质特征并不是相互孤立的,它们与前文中所提到的情感维度空间定义是密切相连的。通过 Pereira 等人的研究表明语音信号的韵律特征与 3 个情感维度(效价维、激活维和控制维)之间具有一定关联性,其中激活维和韵律特征之间具有明显关联,激

活维相近的情感状态具有相似的韵律特征且易混淆。

到目前为止,对情感特征参数的有效提取主要集中在韵律和音质方面,其中以韵律特征为主,而随着研究的深入,越来越多的音质参数也被纳入考虑范围内。前面所提到的特征大多是线性特征,而近几年来各种非线性特征逐渐引起人们重视,其代表性的如 TEO 能量算子。而针对不同民族不同语种对情感表达影响的研究则鲜少见到。此外,由于工作环境的变化,而造成人们不同以往的情绪表达特征的变化也是值得关注的地方。

8 实用语音情感识别研究展望

今后的研究工作可能在情感模型和情感特征方面有较大的发展空间。首先,情感维度空间模型在语音情感识别中的应用还刚刚开始,诸多算法可以与之结合,出现更为合理的情感识别方法。虽然心理学中的"唤醒度-效价度-控制度"三维模型比较流行,但是可以从语音信号的实际特点出发研究更加合适的情感模型。其次,情感特征还有待进一步研究,从声学特征到心理状态的映射是非常困难的,如何构造可靠的情感特征一直是本领域的一个主题。特别是结合跨语言和跨数据库的研究,有利于发掘情感特征中的通用性。

虽然情感计算的研究已经进行了多年,然而情感的科学定义还并不明确。情感可以从进化论得到解释,认为情感是动物在生存斗争中获得的能力,使得动物能够趋利避害。情感还可以从社会心理学的角度得到解释,人类作为群居动物,成员个体之间需要进行有效的沟通,为劳动协作建立关系,而情感则是一种有效的交流手段,体现出个体的意图和心理状态。从这个角度来看,人工智能是不可缺少的情感识别技术,它能够进行复杂意图信息的直接表达和有效传递。

人类语音当中包含的丰富多彩的情感信息,计算机能够理解到何种程度?语音情感识别技术是仅能够模仿一部分的人类情感感知能力,还是有可能超越人类的能力,捕获到人耳亦所无法感知的信息?这些问题值得深思。

从情感的含义上看,既然只有人类和动物才具有情感,那么人类的情感也就通过人类自身得到了界定,人耳所不能感知到的信息,似乎不在语音情感的范畴内。然而,情感的感知通道,并不仅限于人耳听觉。通过内省知觉的方式,说话人自身能够体验到的情感是"体验情感"(Felt emotion),通过人耳听觉感知到的他人的情感,是"听辨情感"(Perceived emotion)。从这个角度考虑,语音情感识别技术,有可能超过人耳的听辨能力,获取到更多的说话人的体验情感的信息。人们在日常生活和工作中无意识地流露出的情感心理状态,能够通过情感计算技术得到准确的测量和分析,在此基础上发展出的技术应用有着广阔的前景。

烦躁情感具有特殊的应用背景,在某些严酷的工作环境中,烦躁是较为常见的、威胁性较大的一种负面情感。保障工作人员的心理状态健康是非常重要的环节。本文中设想在未来可能的长期的载人任务中,对航天员情感和心理状态的监控与干预是一个重要的研究课题。在某些特殊的实际应用项目中,工作人员的心理素质是选拔和训练的一个关键环节,这是由于特殊的环境中会出现诸多的刺激因素,引发负面的心理状态。例如,狭小隔绝的舱体内环境、严重的环境噪声、长时间的睡眠剥夺等因素,都会增加工作人员的心理压力,进而影响任务的顺利完成。因此,本文设想在天地的通信过程中,有必要对航天员的心理健康状况进行检测,在发现潜在的负面情绪威胁的情况下,应该及时地进行心理干预和疏导。在心理学领域,进行心理状态评估的方法,主要是依靠专业心理医师的观察和诊断,而近年来的情感计算技术,则为这个领域提供了客观测量的可能。本文设想,语音情感识别技术可以用于分析载人航天任务中的语音通话,对说话人的情感状态进行自动的、实时的监测。一旦发现烦躁状态出现的迹象,可以及时地进行心理疏导。

参考文献:

[1] Picard R W. Affective computing[M]. Cambridge: MIT Press, 1997.

[2] 赵力. 语音信号处理[M]. 北京:机械工业出版社, 2003.
Zhao Li. Speech signal processing[M]. Beijing: Machinery Industry Press, 2003.

[3] Picard R W. Toward computers that recognize and respond to user emotion[J]. IBM Technical Journal, 2000, 38(2): 705-719.

[4] Scherer K R, Banziger T. Emotional expression in prosody: A review and an agenda for future research [C]// SP2004(Speech Prosody 2004). Nara, Japan: International Speech Communication Association, 2004:355-369.

[5] 赵力, 王治平, 卢韦, 等. 全局和时序结构特征并用的语音信号情感特征识别方法[J]. 自动化学报, 2004,30(3): 423-429.

Zhao Li, Wang Zhiping, Lu Wei, et al. Speech emotional recognition using global and time sequence structure feature[J]. Acta Automatica Sinica, 2004, 30(3): 423-429.

[6] 王治平,赵力,邹采荣. 基于基音参数规整及统计分布模型距离的语音情感识别[J]. 声学学报, 2006,31(1): 28-34.
Wang Zhiping, Zhao Li, Zou Cairong. Emotional speech recognition based on modified parameter and distance of statistical model of pitch [J]. Acta Acustica, 2006,31(1): 28-34.

[7] Arnold M. Emotion and personality[J]. Psychological Aspects, 1960,1:11-116.

[8] Tomkins A S S. The negative affects[J]. Affect, Imagery, Consciousness, 1962,2:111-116.

[9] Murray I, Amott J L. Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion[J]. Journal of the Acoustic Society of America, 1993,93(2): 1097-1108.

[10] Ortony A, Turner T J. What's basic about basic emotions[J]. Psychological Review, 1990, 97(3): 315-331.

[11] Stibbard R M. Vocal expression of emotions in monlaboratory speech: An investigation of the reading/leeds emotion in speech porject annotation data[D]. UK: University of Reading,2001.

[12] Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech[J]. Speech Communication, 2003,40: 5-32.

[13] Taylor G, Fellenz W A, Cowie R, et al. Towards a neural-based theory of emotional dispositions[C]// IMACS/IEEE CSCC. Athens, Greece: IEEE Computer Society, 1999:1-6.

[14] Plutchik R. The multifactor-analytic theory of emotion[J]. Journal of Psychology, 1960,50(1): 153-171.

[15] Ververidis D, Kotropoulos C. A state of the artreview one motional speech databases[C]// Proc 1st Richmedia Conference. Lausanne, Switzerland: IEEE Computer Society, 2003: 109-119.

[16] Douglas-Cowie E, Campbell N, Cowie R, et al. Emotional speech: Towards a new generation of databases[J]. Speech Communication, 2003,40: 33-60.

[17] Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods [J]. Speech Communication, 2006,48: 1162-1181.

[18] Russell J A. Measures of emotion[M]. San Diego, CA, US: Academic Press, 1989.

[19] Cowie R, Douglas-Cowie E. Automatic statistical analysis of the signal and prosodic signs of emotion in speech[C]// Proc ICSLP. Philadelphia, PA, USA: IEEE Signal Processing Society, 1996:1989-1992.

[20] Edgington M. Investigating the limitations of concatenative synthesis[C]// Proc Eurospeech Rhodes. Greece: IEEE Signal Processing Society, 1997: 593-596.

[21] Fernandez R, Picard R W. Modeling drivers' speech under stress [J]. Speech Communication, 2003, 40(1): 145-159.

[22] Fischer K. Annotating emotional language data[R]. Tech. Rep. 236. Germany: University of Hamburg, 1999:111-116.

[23] Yu F, Chang E, Xu Y-Q, et al. Emotion detection from speech to enrich multimedia content[C]// Proc 2nd IEEE Pacific-Rim Conference on Multimedia. Shanghai, China: IEEE Signal Processing Society, 2001:1-6.

[24] Nakatsu R, Solomides A, Tosa N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities[C]// Proc IEEE Int Conf Multimedia Computing and Systems. Floernce, Italy: IEEE Signal Processing Society, 1999: 804-808.

[25] Iida A, Campbell N, Iga S, et al. A speech synthesis system with emotion for assisting communication[C]// Proc ISCA Workshop (ITRW) on Speech and Emotion. Belfast: IEEE Signal Processing Society, 2000:167-172.

[26] Rosenberg A E, Lee C-H, Soong F K. Sub-word unit talker verification using hidden markov models [C]// Proc ICASSP90. New Mexico, USA: IEEE Signal Processing Society, 1990:269-272.

[27] Chasaide A N, Gobl C. Voice quality and the synthesis of affect[J]. Improvements in Speech Synthesis, 2002,25(8): 252-263.

[28] Gobl C, Chasaide A N. Testing affeetive correlates of voice quality through analysis and resynthesis[C]// ISCA Workshop on Speech & Emotion. Northern Ireland: IEEE Signal Processing Society, 2000:1-6.

[29] Kwon O W, Chan K, Hao J, et al. Emotion recognition by speech signals[C]// Proc of Eurospeech. Geneva, Switzerland: IEEE Signal Processing Society, 2003:125-128.

[30] Jianxia C. A summary about emotional speech recognition[C]// The 1st Chinese Conference on Affective Computing and Intelligent Interaction. Beijing: IEEE Signal Processing Society, 2003:11-116.

[31] Tank A E, Kotz S. Accentuation and emotions-two different systems[C]// ISCA Workshop (ITRW) on Speech and Emotion. Newcastle, Northern Ireland: IEEE Signal Processing Society, 2000:1-6.

[32] Gobl C, Chasaide A N. The role of voice quality in communicating emotion, mood and attitude [J]. Speech Communication, 2003,40(1): 189-212.

[33] Tato R, Santos R, Kompe R, et al. Emotion space improves emotion recognition [C] // Proc ICSLP 2002. Denver, Colorado: IEEE Signal Processing Society, 2002: 2029-2032.

[34] Pao Tsang-Long, Chen Yu-Te, Yeh Jun-Heng, et al. Detecting emotions in mandarin speech[J]. Computational Linguistics and Chinese Language Processing, 2005,10(3):347-362.

[35] Ververidis D, Kotropoulos C, Pass J. Automatic emotional speech classification [C] // Proceedings of ICASSP. Montreal, Quebec, Canada: IEEE Signal Processing Society, 2004: 593-596.

[36] Jiang Dan-Ning, Cat Lian-Hong. Speech emotion classificaiton with the combination of statistic features and temporal features[C] // IEEE Intenraitonal Conference on Multimedia and Expo. Taiwan, China: IEEE Computer Society, 2004: 1967-1970.

[37] Audibert N, Auberg V, Rilliard A. Acted vs. spontaneous expressive speech: Perception with inter-individual variability [C] // Proc LREC. Marrakech, Morocco: IEEE Computer Society, 2008: 111-116.

[38] Batliner A, Steidl S, Nth E. Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU aibo emotion corpus[C] // Proc of a Satellite Workshop of LREC. Berlin, Germany: IEEE Computer Society, 2008: 28-31.

[39] Brummer N. Discriminative acoustic language recognition via channel-compensated GMM statistics[C] // ISCA Proc Interspeech. Denver, USA: ISCA, 2009: 1-6.

[40] Busso C, Narayanan S S. Recording audiovisual emotional databases from actors: A closer look[C] // Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International Conference on Language Resources and Evaluation. Amsterdam, Netherland: IEEE Computer Society, 2008: 17-22.

[41] Krajewski J, Kroger B. Using prosodic and spectral characteristics for sleepiness detection[C] // 10th European Conference on Speech Communication and Technology. Geneva, Switzerland: IEEE Computer Society, 2007: 1841-1844.

[42] Yamada T, Hashimoto H, Tosa N. Pattern recognition of emotion with neural network[C] // Industrial Electronics, Control, and Instrumentation, Proceedings of the 1995 IEEE IECON 21st International Conference on. New York, USA: IEEE, 1995, 1: 183-187.

[43] Nwe T L, Foo S W, De Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603-623.

[44] 赵力，钱向民，邹采荣，等. 语音信号中的情感识别研究[J]. 软件学报，2001，12(7)：1050-1055.
Zhao Li, Qian Xiangmin, Zou Cairong. A study on emotional recognition in speech signal[J]. Journal of Software , 2001, 12(7): 1050-1055.

[45] 赵力，将春辉，邹采荣，等. 语音信号中的情感特征分析和识别的研究[J]. 电子学报，2004，32(4)：606-609.
Zhao Li, Jiang Chunhui, Zou Cairong. A study on emotional feature analysis and recognition in speech [J]. Acta Electronica Sinica, 2004, 32(4): 606-609.

[46] 王治平，赵力，邹采荣. 利用模糊熵进行参数有效性分析的语音情感识别[J]. 电路与系统学报，2003，3(8)：109-112.
Wang Zhiping, Zhao Li, Zou Cairong. Emotion recognition of speech using fuzzy entropy effectiveness analysis[J]. Journal of Circuits and Systems, 2003, 3(8): 109-112.

[47] 黄程韦，金赟，赵艳，等. 实用语音情感数据库的设计与研究[J]. 声学技术，2010，29(4)：396-399.
Huang Chengwei, Jin Yun, Zhao Yan, et al. Design and establishment of practical speech emotion database[J]. Technical Acoustics, 2010, 29(4): 396-399.

作者简介：赵力(1958-)，男，教授，博士生导师，研究方向：信号处理等，E-mail：zhaoli@seu.edu.cn；黄程韦(1984-)，男，博士研究生，研究方向：语音信号处理、模式识别等。